공학석사 학위논문

대규모 TV 시청로그 클러스터링을 통한 시청행위 및 시청가구 유형 분석 연구

2017 년 8 월

서울대학교 대학원

융합과학부 디지털정보융합전공

이 태 영

대규모 TV 시청로그 클러스터링을 통한 시청행위 및 시청가구 유형 분석 연구

지도 교수 서 봉 원

이 논문을 공학석사 학위논문으로 제출함

2017 년 6 월

서울대학교 대학원

융합과학부 디지털정보융합전공

이 태 영

이태영의 공학석사 학위논문을 인준함

2017 년 6 월

위 원 장 이 원 종 (인)

부위원장 이 중 식 (인)

위 원 서 봉 원 (인)

초 록

최근에는 더 이상 과거처럼 같은 시간에 온 가족이 모여 앉아 소위 ‘본방사수’를 하는 행동만으로는 TV 시청을 이해할 수 없을 만큼 TV 시청 행태가 매우 복잡해졌다. 다양한 매체와 콘텐츠 공급 서비스들과 상호작용하며 서로 얽히고설킨 복잡한 시청 행동을 보이고 있는 것이다. TV 시청 환경은 콘텐츠 플랫폼 및 디바이스 환경 변화로 인해 과거와 달리 훨씬 예측하기 어려운 복잡한 환경으로 변모하게 되었다.

TV를 둘러싼 환경이 더욱 복잡해지는 상황에서도 TV 시청에 대한 이해는 여전히 중요하게 여겨진다. N-스크린 시청 환경이 보편화되면서 TV에 대한 비중이 하락하고는 있으나, 아직까지는 TV 시청에 많은 시간을 보내고 있고 일상 생활에서의 중요도도 높은 만큼, TV는 여전히 콘텐츠 소비에 있어 중요한 역할을 하고 있기 때문이다. 달라진 환경 속에서도 TV 시청은 여전히 건재한 여가 활동이라 할 수 있으며, 복잡한 환경 속에서 달라진 TV 시청자와 시청 행태에 대한 이해가 더욱 필요한 상황이라는 점을 시사한다.

본 연구에서는 행동을 중심으로 TV 시청을 이해하고자 했던 기존 연구의 확장성 등의 한계점을 극복하고 궁극적으로는 전통적 관점에서 벗어나 다변화된 TV 시청 환경에서의 TV 시청에 대한 이해의 폭을 넓히고자, 디지털 케이블 TV 셋톱박스 로그를 통해 획득한 대규모 TV 시

청 로그를 바탕으로 TV 시청 패턴을 행동 중심으로 유형화하고, 이를 다시 사용자 중심으로 조합하여 해석하는 프레임워크를 제시하였다. 이를 위해 기존의 웹 사용 마이닝 분야에서 사용되었던 세션 클러스터링 기반 유형화 분석 기법을 TV 시청 로그에 적용하였다. 또한 유형화 된 시청 행동과 서비스 해지율 간의 상관관계를 살펴봄으로써 본 연구의 접근 방식이 유효함을 입증하고자 하였다.

제안된 분석 프레임워크를 통해, 본 연구에서는 총 7개의 시청 행동 유형과 이를 통해 조합된 8개의 시청 가구 유형을 도출하였다. 또한 각 시청 가구 유형 그룹 내의 서비스 해지율과 시청 행동 유형 구성비 간의 상관관계 도출을 통해, 본 연구에서 도출한 시청 행동 유형이 서비스 해지를 의미 있게 설명할 수 있다는 것을 확인할 수 있었다.

본 연구를 통해 제안된 분석 프레임워크를 활용하여 행동을 기반으로 시청 패턴을 분석함으로써, 기존의 거시적 맥락에서 이루어진 선행 연구를 확장하여 현재 미디어 환경에서의 TV 시청 행위에 대해 더욱 풍부한 이해를 도울 수 있을 것으로 기대한다.

주요어 : TV 시청 행위 분석, 빅데이터 분석, 행동 기반 분석

학 번 : 2015-26038

목 차

수식 목차

표 목차

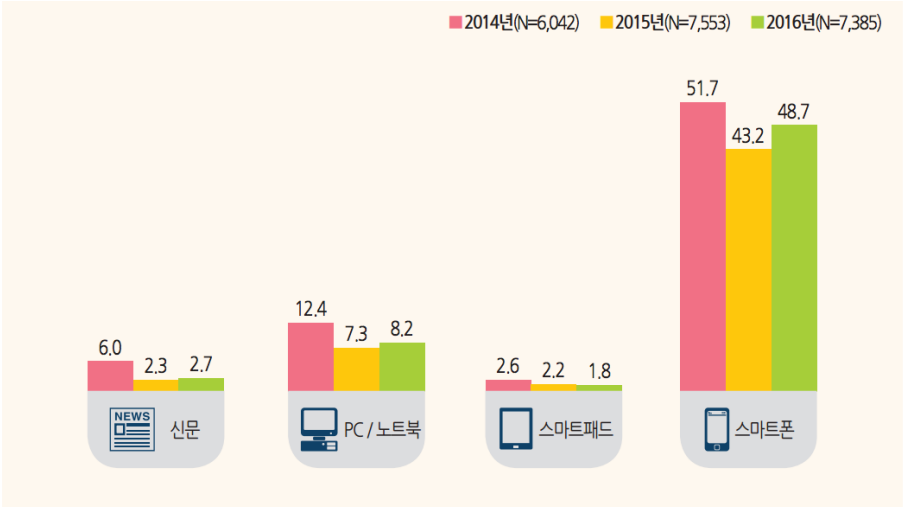
그림 목차

제 1 장 서 론

제 1 절 연구의 배경

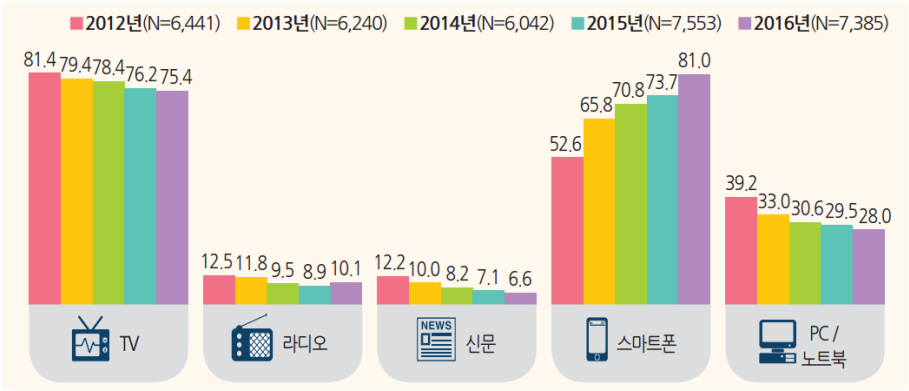
뉴스가 시작되는 저녁 9시에 온 가족이 모여서 지상파 뉴스를 시청하던 모습은 이제 익숙하기보다는 낯선 풍경에 가까워졌다. 오히려 익숙한 풍경은 좀 더 다른 모습이 될 것이다. 사람들은 이제 TV를 배경으로 틀어 놓고 스마트폰을 하기도 하며, 좋아하는 프로그램에 대해서는 VOD를 구매하여 보기도 하고, 심심한 시간에는 TV 채널을 열심히 돌리며 프로그램을 탐색하기도 한다. 한 사람 안에서도 이렇게 다양한 TV 시청 행동과 콘텐츠 소비가 일어나고 있는 것이다. 이처럼 최근에는 더 이상 과거처럼 같은 시간에 온 가족이 모여 앉아 소위 '본방사수'를 하는 행동만으로는 TV 시청을 이해할 수 없을 만큼 TV 시청 행태가 매우 복잡해졌다. 다양한 매체와 콘텐츠 공급 서비스들과 상호작용하며 서로 얽히고설킨 복잡한 시청 행동을 보이고 있는 것이다. 실제로 [그림 1]와 같이 방송통신위원회가 조사한 2016년 방송매체 이용행태조사 결과에서와 같이, TV 시청 중 스마트폰을 이용하는 등 다변화된 시청 행태가 더욱 보편적이다 [1].

TV 시청 환경이 이처럼 변화를 맞이하게 된 이유를 추측하는 것은



[그림 1] – TV 시청 중 타 매체 동시 이용 현황

그리 어렵지 않다. 먼저, 콘텐츠 플랫폼의 다양화 및 융합을 원인으로 들어볼 수 있다. 디지털 기술의 발전에 따라 유·무선통신과 방송 플랫폼이 융합되어 디지털 케이블 TV, IPTV와 같은 다양한 형태의 서비스가 등장하게 된 것이다. 이외에도 푹(Pooq), 티빙(TVing) 등과 같은 OTT(over the top) 서비스와 유튜브(YouTube) 등 동영상 공유 플랫폼에 대한 이용이 확산되고 페이스북 등의 소셜 네트워크 서비스에 의한 콘텐츠 공유가 활발해지는 다매체·다채널 환경이 형성되었다는 점도 이러한 경향을 가속화한다 [2]. 다음으로, TV 시청 행태의 다변화에 콘텐츠 플랫폼의 확장만큼이나 큰 영향을 미친 것으로 스마트폰, 태블릿PC 등이 만드는 N-스크린 환경을 들 수 있다. 방송매체 이용행태 조사나 여타 TV 연구 등에서 TV 시청과 멀티태스킹에 초점을 맞추는 이유도 바로 이 때문이다 [3]. TV를 보는 행동이 다양한 디바이스 사용이라는



[그림 2] – 각 매체를 주 5 일 이상 사용하는 비율

행동의 개입으로 인해 크게 달라졌다는 가정이 전제되어 있기 때문이다. 실제로 TV 시청 시간의 40% 이상이 스마트폰 등의 기기를 멀티태스킹 하는 데 사용하는 것으로 밝혀지기도 했다 [4]. 이처럼 TV 시청 환경은 콘텐츠 플랫폼 및 디바이스 환경 변화로 인해 과거와 달리 훨씬 예측하기 어려운 복잡한 환경으로 변모하게 되었다.

TV를 둘러싼 환경이 더욱 복잡해지는 상황에서도 TV 시청에 대한 이해는 여전히 중요하게 여겨진다. N-스크린 시청 환경이 보편화되면서 TV에 대한 비중이 하락하고는 있으나, 아직까지는 TV 시청에 많은 시간을 보내고 있고 일상 생활에서의 중요도도 높은 만큼, TV는 여전히 콘텐츠 소비에 있어 중요한 역할을 하고 있기 때문이다. 2016년 방송매체 이용행태조사 자료를 살펴보면([그림 2]), TV를 주 5일 이상 사용하는 비율이 2012년 이래로 점차 소폭 감소하는 추세이며 그에 반해 스마트폰의 이용 비율은 급격한 상승세를 보이고 있다 [1]. 그러나 TV를 주 5일

이상 사용하는 비율이 75.4%에 달하고, 하루 평균 TV 이용 시간도 2시간 57분에 이르는 등 여전히 큰 비중을 차지하고 있다. 이는 달라진 환경 속에서도 TV 시청은 여전히 건재한 여가 활동이라 할 수 있으며, 복잡한 환경 속에서 달라진 TV 시청자와 시청 행태에 대한 이해가 더욱 필요한 상황이라는 점을 시사한다.

TV 시청 행태 및 시청자 프로파일은 경영학 [5], 컴퓨터공학 [6], 산업공학 [7] 등 다양한 학문 분야의 관심사가 되어왔다. 이러한 이종의 학문 분야에서 공통적으로 나타나는 흐름 중 하나는 TV 시청 패턴을 인구통계학적 요소 및 교육·소득수준 요소와 결합하는 것이었다 [8, 9, 10]. 거시적인 관점에서 시청 패턴과 시청자·시청 가구 사이의 관계를 찾고자 했던 것이다. 이러한 연구들은 각 시청자의 시청 패턴이나 매체 간 이용 성향이 어떤 요인과 연관되어 있는지에 대해, 특징적인 개별 시청자 집단의 수준에서 시청자와 시청 행위 간의 관계를 입증하고자 했다. 문제는 인구통계학적 요소나 취향과 같은 특정 요인이 TV 시청 패턴을 결정한다는 전통적인 방식과 관점으로는 여러 요인으로 인해 한 사람·한 가구 내에서도 다양한 시청 행동이 벌어지는 현재 상황을 설명하기가 매우 어렵다는 점이다.

기존의 TV 시청 행동을 이해하려는 이같은 접근들이 복잡한 시청 환경 속에서 나타나는 다양한 차원의 행동 그 자체에는 관심을 기울이지 않는다는 한계를 지닌다. 인구통계학적 요인 분석이나 취향 정의를 통한

접근이 TV 시청 행동을 이해하는 데 다소 도움을 준 것은 사실이지만, 다양화된 TV 시청 행동을 개별적으로 이해하기는 어렵다. 복잡한 TV 시청 환경처럼 다양한 요인이 영향을 주며, 개별 사용자의 기분과 같이 특정 순간의 선택에 영향을 주는 요인이 있는 행동의 경우에는 한 사람을 둘러싼 속성보다는, 개별 행동 중심의 분류를 통한 이해가 더욱 효과적일 수 있다 [11]. 예를 들어, 성별과 직업을 통해 TV 시청 패턴을 분석할 경우, '여성들은 낮시간대에 TV를 많이 본다', '회사원의 경우 주말 TV 시청량이 많다'는 식의 결과에 도달할 수 있다. 이러한 방식의 전통적 분석에서는 최초로 설정한 요인인 성별, 직업을 중심으로만 결론을 해석할 수밖에 없다. 하지만 다변화된 환경 속에서만 나타나는 TV 시청 행태를 행동 중심으로 모델링할 경우, '배경적으로 TV를 시청하는 행동', '특정 TV 프로그램을 몰아보는 행동' 등을 먼저 발견하고, 이것들이 기존에 밝혀진 여러 요인들과 어떠한 관계가 있는지 살펴볼 수 있어 변화된 TV 시청을 이해하는 폭을 넓힐 수 있다.

이 같은 흐름에 따라 최근의 TV 시청 이해에 대한 연구도 행동 중심으로 사용자를 이해하고자 하는 시도를 지속해왔다. Vanattenhoven 등 [12]은 TV 시청을 둘러싼 다양한 행동을 유형화하고자 하였다. 그들은 개별적 시청 패턴에 관심을 기울이기보다 시청시간대, 개별 이벤트, 함께 시청하는 사람, 시청 콘텐츠 등을 중심으로 행동을 분류하는 작업을 진행하였다. TV 시청에 대한 사용자 유형의 분류가 아닌 행동 유형을

밝히려고 시도했던 것이다. 또한 Trouleau 등 [13]의 경우에는 특정하게 유형화 된 TV 시청 패턴으로서 'VOD 몰아보기'를 규정하고, 이러한 행동 유형 내부의 메커니즘을 규명하고자 했다. 이러한 최근의 연구 경향은 개별 사용자 중심으로 시청자 유형을 분류하는 기존의 접근의 한계를 극복하고자 하는 것이며, 정의된 행동 유형 중심으로 사용자를 설명하는 것이 더욱 효과적이라는 점을 시사한다.

행동 중심으로 TV 시청 행태를 이해하는 기존의 연구들은 관점의 전환과 사용자에 대한 깊은 이해라는 토대를 마련했지만, 두 가지 문제점을 극복하지는 못했다. 첫째로, 행동 중심의 TV 시청 행태 분석 연구는 데이터 수집 및 분석에 드는 상당한 비용 때문에 그 규모가 매우 제한적이며, 이것이 결과적으로는 새로운 결과를 발굴하는 데 어려움을 준다는 점이다. 인터뷰나 관찰을 통해 진행되는 연구에서는 10명 내외를 유효한 분석 대상으로 상정하고 있으며 [14], 설문 연구의 경우에도 몇 만 건 이상이 되는 대규모의 데이터를 수집하기는 어렵다. 이러한 데이터 자체의 한계는 새로운 유형의 행동을 발견하거나 분류하기 어렵다는 문제로 이어지게 된다. 둘째로는, 유형화 된 행동을 어떤 방식으로 활용할 수 있는지에 대한 구체적인 가이드라인을 제시하지 못했다는 점이다. 도출된 행동 유형이 사용자를 이해하는 데 어떻게 도움을 줄 수 있을지에 대한 구체적인 시사점(implications)을 제공할 수 있어야 향후 TV 시청에 대한 연구, 산업적 전략 등에 활용이 가능함에도, 앞선 연구들은

이를 제시하지 못했던 것이다.

한편, 변화하는 TV 시청 환경에서 시청 행태를 이해하기 위한 연구의 흐름으로 다채널 환경에서 시청자가 채널을 선택하는 행위의 능동성과 수동성에 대한 논의 및 채널 소비의 다양성에 관한 논의가 이루어졌다. 지상파 채널이 중심이 되는 채널 환경에서는 구조적 요인으로 인해 TV 시청 행위가 수동적인 모습을 보인다는 시각이 주를 이뤘으나, 현재는 다양한 채널에서 특정 장르의 프로그램을 상시 제공하는 전문 채널이 등장하는 등 선택의 폭이 매우 다양해 졌으므로, TV 시청 행위가 개인의 선호를 적극 반영한 능동적 행위라는 시각이 주를 이루고 있다 [15, 16]. 그러나 채널 공급의 다양성이 반드시 소비의 다양성을 불러일으키지는 않는다는 현상을 채널 레퍼토리 및 시청자 분극화 경향 연구가 지적하였다 [17]. 연구들은 이러한 경향을 인구통계학적 요인 등에 따라 나타나는 차이를 통해 설명하였다. 이러한 연구들을 통해, 다채널 환경에서 TV 시청 행위를 이해하는 데에 있어 능동성과 수동성, 그리고 소비의 다양성이 매우 중요한 역할을 한다는 것을 알 수 있다.

본 연구에서는 앞에서 제시한, 행동을 중심으로 TV 시청을 이해하고자 했던 기존 연구의 한계점을 극복하고 궁극적으로는 전통적 관점에서 벗어나 다변화된 TV 시청 환경에서의 TV 시청에 대한 이해의 폭을 넓히고자 한다. 이를 위해 본 연구에서는 (1) 디지털 케이블 TV를 통해 획득한 2억 개의 대규모 TV 시청 로그를 바탕으로 (2) TV 시청 패턴을

행동 중심으로 유형화한 후, (3) 이를 다시 사용자 중심으로 조합하여 해석하는 프레임워크를 제시하고자 한다. 특히 TV 시청 패턴을 행동 중심으로 유형화하는 데에 있어, 시청자의 능동성·수동성을 고려하여 채널 전환 및 선택이 적극적으로 일어나는지 혹은 소극적인 채널 이용을 보이는지를 살펴보고, 채널 이용의 다양성 또한 고려하고자 한다.

제시되는 프레임워크를 통해, 본 연구에서는 기존의 웹 사용 마이닝(Web Usage Mining) 분야에서 주로 사용되는 클릭스트림(clickstream) 및 세션(session) 클러스터링을 통한 유형화 방안을 TV 시청 로그 분석에 도입하고자 한다. 인터넷 환경에서 사용자의 인터넷 브라우징을 이해하고자, 사용자가 발생시킨 클릭스트림을 세션화해 사용자의 인터넷 브라우징 행동을 모사하고 이를 바탕으로 사용자 또는 행동을 세분화하고자 하는 연구들이 다수 이루어졌다 [18, 19]. TV 시청은 웹 브라우징과 비교해 봤을 때 겉보기의 형태에서는 차이가 있으나, 소비자가 자신이 원하는 콘텐츠를 탐색하고 적합한 콘텐츠를 발견하여 소비한다는 행동 패턴 면에서 연결고리가 있다. 물론 TV의 프로그램이 대략 한 시간의 방영시간을 가져 한 콘텐츠를 소비하는 데에 드는 시간에 차이가 있으나, 본 연구에서는 두 영역에서 나타나는 행동 패턴의 유사성에 착안하여 세션 클러스터링을 통한 시청 패턴 유형화의 가능성을 확인해보고자 한다.

또한, 본 연구에서는 특정 디지털 케이블 TV 서비스의 시청 로그를 바탕으로 행동 중심 시청 행위 유형화 및 시청 가구 유형화를 수행함에

있어, 이러한 방식의 유형화에 산업적 시사점이 있음을 검토하기 위해 시청 행동과 서비스 해지 간의 연관성을 입증해보고자 한다. 한 달 간의 시청 로그를 바탕으로 각 시청 가구 유형이 다음 달에 보이는 해지율과 시청 행위 유형 간의 상관관계를 도출함으로써, 어떠한 시청 행위 유형이 서비스 해지에 영향을 미치는지를 파악해보고자 한다.

제 2 절 연구의 목표

본 연구는 디지털 케이블 TV 서비스의 셋톱박스를 통한 시청 로그에 기반하여 TV 시청 가구를 모델링하고자 한다. 특히 TV 시청 가구들이 어떤 패턴의 시청 행위를 일으키는지 보다 구체적으로 이해하기 위해, TV 시청 가구를 모델링함에 있어 두 단계의 접근법을 제시하고자 한다. TV 시청 가구의 콘텐츠 이용은 곧 TV 시청 행위의 집합이므로, 먼저 TV 시청 행위 자체를 유형화하여 분석한 후 이를 바탕으로 TV 시청 가구를 모델링하고자 한다.

본 연구의 첫 번째 목표는 다음과 같다.

연구 목표 1: 현재의 디지털 케이블 TV 시청 가구들이 어떤 특성의 시청 행위를 가지고 있는지 알아보기 위하여, TV 시청 가구의 행동적 특성을 유형화하는 방안을 제시한다.

먼저 TV 시청 행위를 유형화하기 위해 각 시청 가구의 셋톱박스 이용 로그를 조합해 TV 시청이 시작될 때부터 종료될 때 까지를 한 단위로 하는 세션(session)으로 변환한다. 여기서 TV 시청 행위는 문제를 단순화하기 위해 실시간 TV 채널 시청 행위로 한정해, 시청자가 채널을 변경하고 프로그램을 시청하는 것을 의미한다. TV 시청 행위의 특성은 세션의 길이, 시청한 채널의 수, 채널 당 머무른 시간 등 세션에서 나타나는 서비스 이용 특징을 의미한다. 세션으로 변환된 시청 로그에서 서

비스 이용 특징을 잘 표현할 수 있는 특질을 추출하여 세션을 모델링하고, 모델링 된 세션을 클러스터링 기법으로 그룹 짓고 각 클러스터의 특징을 분석하여 유형화한다.

이어서 본 연구의 두 번째 목표는 다음과 같다.

연구 목표 2: TV 시청 가구들이 어떤 시청 행위 유형과 서비스 이용 특성의 조합으로 이루어져 있는지 알아보기 위하여, 시청 행위 유형을 바탕으로 디지털 케이블 TV 시청 가구를 모델링하는 방안을 제시한다.

각 시청 가구를 모델링한다는 것은 곧 시청 가구의 특징을 잘 나타내는 특질을 찾아내는 것이라 할 수 있다. 본 연구에서는 이에 대해 각 시청 가구가 발생시킨 세션의 유형을 구분하여, 각 세션 유형의 비율을 시청 가구의 특질로 삼는 행위 기반 모델링을 제안하고자 한다. 세션 유형의 비율로 모델링 된 시청 가구를 클러스터링하여, 각 시청 가구 클러스터의 특징을 분석하여 시청 가구의 유형을 분류한다.

다음으로 본 연구의 세 번째 목표는 다음과 같다.

연구 목표 3: 시청 행위 유형을 바탕으로 묘사된 각 시청 가구 유형의 서비스 해지율을 통해, 서비스 해지와 연관성이 큰 시청 행위 유형에 대해 파악한다.

앞의 프레임워크를 거쳐 도출된 각 시청 가구 유형은 각 시청 행위 유형의 구성비로 묘사되어 있다. 본 연구에서는 한 달 간의 시청 로그 분석을 통해 구성된 각 시청 가구 유형이 전체 가입기간에 걸쳐 비슷한 시청 행위 패턴을 유지한다고 가정하고, 도출된 시청 가구 유형이 다음 달에 보이는 서비스 해지율과 각 시청 행위 유형의 구성비 간의 상관관계를 계산하여, 시청 행위 유형과 서비스 해지 간의 연관성을 파악해보고자 한다. 이를 통해 본 연구에서 제안하는 시청 행위 및 시청 가구 분석 프레임워크가 TV 시청 패턴에 대한 이해에서 그치는 것이 아닌, 실제로 산업적 이용가치가 있고 시사점을 줄 수 있다는 것을 입증해보고자 한다. 특히, 본 연구에서 사용한 데이터는 디지털 케이블 TV 서비스의 특정 지역 전체 가입자를 대상으로 한 실제 데이터이므로, 연구의 결과에 따라 실제 서비스에 시사점을 곧바로 고려하고 적용할 수 있다는 장점이 있다.

제 2 장 선행연구

제 1 절 이론적 배경

이 절에서는 TV 시청 행위에 관해 이루어졌던 기존의 연구와 함께, 본 연구에서 방법론으로 제시한 클릭스트림 또는 세션 유형화에 대해 온라인 사이트 사용자 연구 분야에서 이루어졌던 연구에 대해 살펴본다. 먼저 다채널 환경에서의 TV 시청 행위 연구에 대해 살펴보고, TV 시청 행위에 있어서 맥락적 요소를 통해 시청 행위를 이해하고자 한 연구에 대해 짚어본 다음, 클릭스트림 및 세션 유형화에 따른 사용자 세분화에 대한 연구를 살펴본다.

1.1. 다채널 환경에서의 TV 시청패턴 연구

1.1.1. TV 시청 행위의 능동성과 수동성

케이블 TV의 등장으로 시청 가능한 채널 수가 늘어남에 따라, 과거 지상파 방송 환경에서 시청자들이 최대 다수에게 두루 소구하는 프로그램에만 만족해야 했던 것과 달리 시청자가 직접 원하는 채널을 능동적으로 선택할 수 있는 환경이 조성되었다 [20]. 과거에는 TV 프로그램을 선택하는 행위가 시청자의 필요에 의해 능동적으로 이루어지기 보다는 해당 시점에서의 가용성에 좌우된다는 의견이 주를 이루었다 [21]. 현재

와 비교해 다소 경직되어 있던 콘텐츠 소비 환경으로 인해, 시청 가용성 및 프로그램 장르, 선택 가능한 프로그램의 수 등의 구조적 요인이 시청자들의 프로그램 선택에 대한 예측을 가능하게 했다 [22].

그러나 다채널 환경이 조성됨에 따라 시청자가 선호하는 프로그램이 대중적이지 않더라도 해당 장르의 프로그램을 공급하는 틈새 채널 [23]을 선택할 수 있게 되면서, 시청자들의 능동적 시청 행위가 기본 전제가 되는 이용과 충족 이론이 다채널 환경에서의 시청 행위 연구에 있어 부각되었다 [16]. 즉, 시청자가 더 이상 수동적 방관자가 아닌 능동적 결정권자이며, 수용자가 아닌 이용자임이 재조명되었다 [15].

공급되는 채널 수의 변화와 시청 행위의 능동성 사이의 관계를 파악하기 위해, 배진아 [24]는 지상파 채널만을 제공하는 시청 환경과 케이블 TV 시청 사이의 시청 패턴 유사성을 바탕으로 다채널 TV 시청의 수동성과 능동성을 비교하는 연구를 진행하였다. 이 연구에서는 다채널 TV 시청 환경에서 기존의 지상파 중심 시청 환경과 비슷한 시청 패턴이 나타나는지의 여부를 통해 시청 환경의 변화가 능동성을 불러일으키는지에 대해 조사하였다. 그 결과 지상파 환경과 다채널 환경에서 모두 장르를 선택하는 데에 있어 시청 행태의 유사성이 발견되었고, 이에 따라 케이블 TV를 통한 다채널화가 능동적 시청 행위를 불러일으키지 못하고 국내 방송 환경에서는 수동적으로 TV를 시청하는 경향이 큼을 밝혔다.

특히 [24]에서는 지상파 환경에서 다채널 환경으로 전이되는 과정에서 시청 행위의 능동성에 관해 논하면서, 능동성을 특정 장르나 프로그램에 대한 선호가 적극적으로 반영되는 시청 행위라 규정하였다. 지상파 환경과 다채널 환경의 가장 큰 차이점은 TV 시청 상황에서 가능한 대안의 수인데, 대안의 수가 기존에 비해 수십 배로 확장된 다채널 환경에서 시청자들의 채널 및 장르 선택 행위는 기존의 그것과 큰 차이가 있다고 볼 수 있다. 따라서, 다채널 환경에서 케이블 채널을 주로 시청하여 장르 추종적 경향이 큰 시청 행위일수록 능동적이며, 그렇지 않고 지상파 또는 종합편성채널 등을 주로 시청하여 장르 추종적 경향이 작은 시청 행위를 수동적으로 정의하였다.

본 연구에서는 [24]의 능동성에 대한 정의를 바탕으로, 유형화 된 시청 가구의 시청 행위를 분석하여 시청 행위가 주로 일어나는 채널이 케이블에 한정될수록 능동적이며, 지상파로 수렴할수록 수동적인 시청 행위를 보인다고 규정하였다. 이를 통해 각 시청 가구가 보이는 능동성을 파악할 수 있다. 또한 본 연구에서는 곧바로 시청 가구의 채널 이용 현황을 분석하여 능동적 시청 행위의 정도를 보는 것이 아닌, 단계를 나누어 우선 시청 행위 자체의 유형을 파악하고 시청 가구를 이 시청 행위 유형의 조합으로 해석하므로, 어떠한 시청 행위 유형이 조합되어 능동성 또는 수동성을 불러일으키는지에 대해 알 수 있어 다채널 환경에서의 시청 행위에 대해 보다 구체적으로 이해할 수 있다는 점에서 기존의 관련

연구와 차별점이 있다고 할 수 있다.

1.1.2. TV 콘텐츠 소비의 다양성

공급되는 채널의 수가 확대되어 시청자들의 다양한 선호를 충족시키고 선택권이 늘어나게 되었지만, 이러한 공급의 다양성이 반드시 소비의 다양성을 초래하지 않는다는 사실이 밝혀졌다. 미디어 정책입안자들은 콘텐츠 공급원 및 내용의 다양성이 소비의 다양성을 불러일으킬 것이라 예측하였는데 [25], 예상과 달리 채널이 증가하여도 시청자들이 주로 소비하는 채널의 수는 거의 증가하지 않았다. 시청자들은 주로 10~15개 내외의 채널에 머물러 있는 현상을 보였는데, 이러한 한정적인 채널 이용 패턴을 가리켜 채널 레퍼토리 [26]라 한다.

채널 레퍼토리를 구체적으로 살펴보면 대부분 지상파와 인기 채널을 공유하는데, 이에 반해 비인기 채널에서는 시청자들마다 서로 다른 장르를 선호하는 경향인 분극화 소비 행태가 밝혀졌다 [27, 17, 28, 29]. [17]에서는 지상파와 케이블 TV의 교양, 다큐, 드라마 및 영화 채널이 주로 공유되는 대집단 채널이며, 분극화가 이루어지는 소집단 채널은 해외 위성, 외국어 방송, 국내 위성 방송, 케이블 TV 종교 채널 등임을 밝혔다. 또한 인구통계학적 요인을 통해 차별화된 시청패턴을 보이는 원인을 분석하였는데, 분극화 요인으로는 성별, 연령대, 교육수준, 소득수준, 그리고 직업의 차이가 나타났다.

마찬가지로 다채널 환경에서의 TV 시청 행위에 대해 이해하고자 하는 본 연구 또한, 앞서 살펴본 TV 시청 행위의 콘텐츠 소비 다양성을 고려하고자 한다. 선행 연구에서 밝혀졌듯이 최근의 TV 시청자는 다채널 환경에서 개인의 선호 요인에 따라 능동적으로 채널을 시청하며 또한 채널 레퍼토리를 형성하는데, 기존 연구에서는 이를 각 개인 단위의 채널 시청 현황과 채널 레퍼토리, 그리고 이를 설명하기 위한 인구통계학적 요인에 집중하여 실제 시청자들의 행동이 어떤 방식으로 이루어지는지에 대해서는 간과한 부분이 있다고 볼 수 있다. 본 연구에서는 이러한 배경 하에서, 구체적으로 각 시청 가구가 시청을 시작하고 종료할 때까지 채널을 선택하고 시청함에 있어 각 시청 가구의 행동이 얼마나 다양한 채널 소비로 이루어지는지 알아보고자 한다. 이를 위해 기존의 연구에서 밝힌 채널 선택과 장르 선택의 다양성을 각 시청 행위에 대해 묘사하고, 이를 기반으로 시청 행위를 유형화하여 시청 패턴을 이해해보고자 한다.

1.2. TV 시청의 맥락에 대한 연구

이 절에서는 가정에서 일어나는 TV 시청 행위와 그 맥락에 대한 연구를 검토한다. 일반적인 경우 이러한 연구는 현장에서 얻어진 정보를 통해 정성적으로 수행되어, 자연스러운 TV 시청에 대한 풍부한 이해를 돕는다.

Barkhuus [30]는 디지털 비디오 레코더를 사용하거나 인터넷을 통해 시청하는 등 새롭게 등장한 TV 시청 환경의 얼리어답터에 대해 심층 인터뷰를 진행했다. 이 민속지 연구는 개인 시청자 및 부부를 대상으로 이루어졌다. 일반적으로 이들은 매일 저녁마다 가족 구성원마다 2~3시간 가량 TV를 시청하며, 이러한 시청 활동을 편안하고 저렴한 활동으로 묘사하였다. 이 시간동안 주로 TV가 켜져 있기는 하지만, TV가 켜져 있는 시간 내내 시청 활동이 일어나지는 않는다. 즉, 이를 통해 Lull [31]이 설명하였듯이 TV의 전통적인 역할인 배경 잡음, 동반자 관계, 혹은 엔터테인먼트의 형태로 TV를 환경적 요소로서 이용하는 현상이 확인되었다.

Saxbe 등 [32]은 30가구를 대상으로 10분 간격으로 가정 내 활동을 기록하여, TV 시청 행위가 어떤 가족 구성원이 언제, 집 안의 어떤 장소에서 일어나는지에 대한 패턴을 연구하였다. 연구에 따르면 전체 TV 시청 행위 중 36%가 개인의 방에서 이루어졌으며, 나머지 시청 시간 중 전체 가족 17%, 자녀 15%, 어머니와 자녀 13.5%, 아버지와 자녀 12.5%, 부모 6%의 구성비로 시청이 이루어짐을 확인하였다. 물론 각 비율은 가구 구성 및 국가에 따라 달라질 수 있으나, TV 시청 행위가 이루어지는 대부분의 시간에 여러 가족 구성원이 참여하며 가구원의 구성이 달라질 수 있음을 밝혔다.

Mercer 등 [33]은 비디오 기반 콘텐츠 사용자 11명에 대한 실지

연구를 통해 콘텐츠 시청 상황의 맥락을 파악하고자 하였다. 이 연구에서는 1인 시청과 공유 시청, 공공장소와 개인 공간, 그리고 시간적 특성에 따라 네 가지 전형적인 콘텐츠 시청 상황을 구별하였다.

TV 시청의 맥락에 관한 연구는 정성적인 방식 뿐만 아니라 정량적으로도 이루어졌는데, Abreu 등 [34]은 새로운 환경에서의 TV 시청 행동에 대해 설문 조사를 실시하였다. 연구 결과에 따르면 주로 저녁 시간에 TV 시청 행위가 많이 발생하며, 아침 및 오후 시간대에는 시청량이 줄어드는 경향을 보였다. 그러나 주말에는 주중에 비해 모든 부분에서 더 많은 TV 시청 행위가 일어났다. 설문에 따르면 가족과 함께 TV를 시청하는 시간이 전체의 52%, 혼자 시청하는 시간이 42% 등으로 나타났는데, TV 시청 대부분이 여러 사람과 함께 일어난다는 사실을 확인하였다. 또한 시청자들이 시청을 결정하는 요소를 중요도 순서대로 나열하였는데, 이는 프로그램의 장르, 그 순간의 심적 상태, 혼자 또는 여러 사람인 상태, 그리고 여유 시간 순으로 나타났다.

Chaney 등 [35]은 4백만 개의 가구 시청 로그를 사용하여 데이터 기반 분석을 수행하였다. 이 연구에서는 함께 시청하는 시간의 80%가 구성원이 두 사람인 경우임을 밝혔으며, 구성원의 수가 늘어날수록 시청한 프로그램 수가 현저하게 줄어든다는 점을 알아냈다. 또한 연령과 성별에 따라 장르별 시청 분포가 달라짐을 보였는데, 뉴스의 경우 남성과 여성 모두 연령이 높아짐에 따라 시청량이 증가하였으나, 스포츠의 경우

여성은 모든 연령층에 걸쳐 시청량이 매우 낮은 반면 남성은 20세에서 30세 사이에서 높은 시청량을 보였다. 각 성별에 따라 여성은 토크쇼와 드라마, 음악 장르에 대한 시청량이 높은 반면, 남성은 애니메이션, 다큐멘터리 및 스포츠 장르를 선호하였다. 또한 그룹 시청에서도 장르에 따른 차이가 드러났는데, 퀴즈쇼와 드라마, 스포츠 장르가 전체 그룹 시청의 1/3을 차지하였으며, 반면 음악, 뉴스 및 정치 관련 프로그램은 그룹 시청의 20%를 차지하였다.

위와 같은 연구를 통해 각 가정에서 TV를 시청하는 방법을 결정하는 다양한 요인 및 맥락에 대해 폭넓은 이해가 가능하다. 본 연구에서는 TV 외적인 맥락 및 인구통계학적 요인을 통해 TV 시청 행위를 설명하는 대신 TV 시청 행위 자체에서 행동적 특성을 파악하고 시청 흐름에 대해 이해해보고자 한다. 본 연구에서 제안하는 방법론 및 결과를 토대로 TV 시청 행위의 구체적인 패턴에 대해 파악함과 동시에 TV 외적 요소와 시청 행위 간의 연구결과를 종합한다면, TV 시청 행위에 대해 더욱 폭넓으며 세밀한 이해를 도울 수 있을 것이다.

1.3. 클릭스트림 유형화 및 분석 연구

사용자의 행동을 잘 이해하기 위한 직관적인 솔루션은 잘 설계된 사용자 연구를 통해 사용자가 시스템을 사용하는 방법을 조사하는 것이다. 그러나 세부적인 사용자 연구는 규모가 제한적일 수밖에 없는데, 이

와 비교하여 클릭스트림 분석은 대규모 사용자 집단으로 확장이 가능하며 이전에 알려지지 않은 행위에 대해 추가적으로 식별이 가능하다는 장점이 있다. 특히 클러스터링 기법은 사용자나 행동에 대한 범주나 레이블 등 사전 지식에 의존하지 않으므로, 알려진 것과 알려지지 않은 행동 패턴을 모두 포착할 수 있다.

사용자 행동을 이해하는 것은 온라인 서비스의 설계 및 운영에 있어 중요한데, 최근의 연구들은 네트워크 트래픽을 분석하여 온라인 사용자의 브라우징 습관을 이해하고자 하였다. Adar 등 [36]은 사용자들이 웹 콘텐츠를 재방문하는 방식을 유형화하기 위해 약 60만 명에 대한 5주 간의 웹 상호작용 로그를 분석하였다. 또한 관찰된 재방문이 어떤 의도를 가졌는지 파악하기 위해 추가 설문조사를 통해 결과를 보완하였다. 이를 통해 행동이나 내용, 또는 구조적 특징을 가진 네 가지의 재방문 패턴을 도출하였다. Obendorf 등 [37] 또한 웹페이지 재방문 유형을 식별하기 위해 장기간의 클릭스트림에 대해 연구하였다. 이 연구를 통해 각 개인의 브라우징 방식이 다양하게 다르며, 특히 사용자 행동 로그 및 인터뷰를 토대로 역추적 또는 실행 취소의 성격을 가진 단기간의 재방문, 재사용의 성격을 지닌 중기 재방문, 그리고 재발견의 성격을 지닌 장기 재방문을 구별하였다.

온라인 소셜 네트워크에서의 악성 계정 탐지를 위해서도 사용자 행동 모델을 도출하기 위해 클릭스트림 연구가 이루어졌는데, Wang 등

[19]의 연구에서는 서버 측의 클릭스트림을 모델링하여 가짜 신원을 탐지하기 위한 시스템을 구축하였다. 이 연구에서는 클릭스트림 시퀀스 간의 거리를 나타낸 유사성 그래프를 분할하여 유사한 사용자의 클릭스트림을 행동 클러스터로 그룹화하였다. 그리고 이 시스템을 22만 명이 사용 중인 중국의 소셜 네트워크 Renren의 실제 및 악성 계정 1만 6천 명에 대한 클릭스트림에 사용하여, 악성 계정의 클릭스트림 추적에서 매우 높은 탐지 정확도를 보인다는 것을 밝혔다.

웹 데이터에서 중요한 사용 패턴을 발견하고 추출하기 위한 연구 또한 이루어졌는데, Lu 등 [18]은 웹 사용 마이닝(Web Usage Mining)을 통해 중요한 사용자 브라우징 경로를 식별하였다. 이 연구에서는 웹 클릭스트림을 세션화하고 특질을 추출하여 세션을 클러스터링한 후, 클러스터링된 세션 클러스터들을 개념을 기반으로 추상화하여 중요 사용 패턴을 생성하기 위한 파이프라인 처리 단계를 제안하였다.

웹 기반 활동이 증가함에 따라 사용자가 웹 페이지를 방문할 때 사용자의 다음 요청을 예측하는 것이 중요한 문제로 대두되었는데, [38]에서는 기존에 널리 사용되던 마르코프 모델이 사용자의 전체 행동을 포착하기 어렵다는 점과 순차 패턴 마이닝 기법이 빈번한 시퀀스만을 고려하여 순차 패턴이 아닌 행동을 예측하기 어렵다는 한계를 지적하며, 세션의 페이지 순서 정보에 따른 유사성을 기반으로 사용자 세션을 클러스터링하고 결과 스트림을 클릭스트림 트리로 나타내는 새로운 모델을 제안

하였다.

본 연구에서는 웹 기반 데이터에서 클릭스트림 또는 세션을 유형화하기 위한 분석기법을 차용하여, TV의 시청 로그에 적용하고자 한다. 본질적으로 웹의 클릭스트림과 TV의 채널 시청 로그는 사용자가 여러 옵션에 대해 탐색하고 그 중 자신에게 적합한 콘텐츠를 선택한다는 점에서 비슷하다. 따라서 클릭스트림을 세션화하고 세션의 유형을 나누는 것과 마찬가지로, TV 채널 시청 로그를 세션화하고 각 세션의 특질을 추출하여 추상화한 뒤, 세션 간의 유사성을 측정하여 클러스터링하는 방식으로 세션을 유형화하고자 한다. 또한 유형화 된 세션을 각 시청 가구가 생성하는 비율과 세션 유형의 흐름을 통해, 대규모 TV 시청 가구의 행동 패턴을 이해해보고자 한다.

제 2 절 기술적 배경

이 절에서는 본 연구에서 활용한 기술적 요소들에 대해 설명하고자 한다. 먼저 대규모 데이터 처리 및 분석에 사용한 빅데이터 분석 프레임워크에 대해 살펴보고, 시청 행위와 시청 가구의 유형화를 위해 사용한 K-Means 클러스터링에 대해 살펴본다.

2.1. 빅데이터 분석 프레임워크

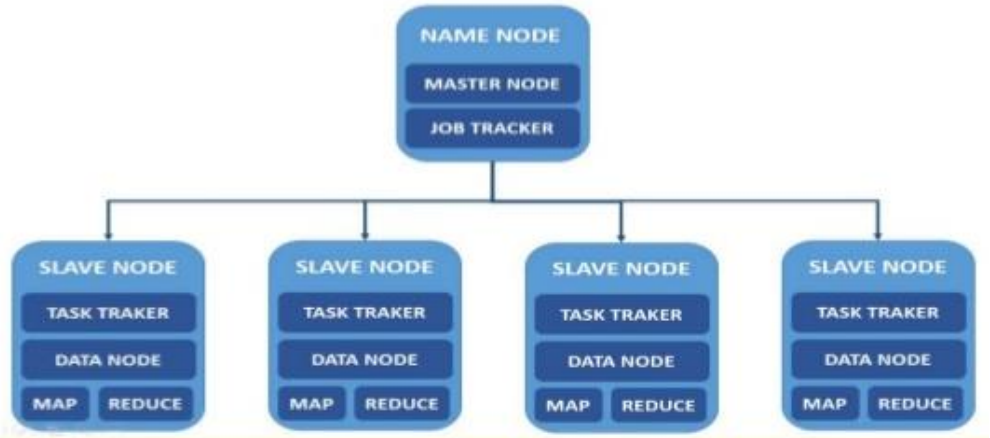
과거의 데이터 분석 기술은 대부분 한 대의 컴퓨터에서 데이터를 파일 시스템, 데이터베이스 또는 메모리 상에 저장하고 이를 기반으로 알고리즘을 실행하는 구조로 이루어졌다. 일반적인 분석 도구들은 파일 시스템의 데이터를 메모리에 적재해 이용하는 방식으로 작동했기 때문에 대용량의 데이터를 처리하기 위해서는 메모리를 증설해야 했고, 이때문에 처리할 수 있는 데이터에 한계가 있었다. 각종 데이터베이스 시스템이 도입되며 처리할 수 있는 데이터의 규모가 증가하기 시작했으나, 이 또한 하나의 머신에 최적화된 형태로 구성되었다. 따라서 증가하는 데이터의 규모에 맞춰 데이터 분석을 수행하기 위해서는, 단일 머신의 중앙처리장치 성능을 높이고 메모리와 디스크의 용량을 늘리는 스케일-업(scale-up) 방식의 규모 확장이 다수를 차지했다.

단일 머신이 감당하기 힘든 대규모의 데이터를 저장하고 현실적인

시간 안에 분석을 수행하기 위해, 구글(Google)에서는 대용량의 데이터를 효과적으로 분산 저장할 수 있는 파일 시스템 기술인 GFS(Google File System) [39]와 알고리즘을 분산 수행하기 위한 프로그래밍 모델인 MapReduce [40]를 확보하여 적극적으로 이용하였다. 이 기술을 논문으로 공개함으로써, 이를 기반으로 다양한 대용량 분산 파일 시스템 및 대규모 분산 처리 프레임워크가 오픈소스 형태로 등장하게 되었다.

여러 오픈소스 분산 파일 시스템 및 분산 처리 프레임워크 중, 야후(Yahoo!)에서 공개한 Java 기반의 아파치 하둡(Apache Hadoop; High-Availability Distributed Object-Oriented Platform)이 가장 커다란 오픈소스 생태계를 이뤄 분산 처리 프레임워크의 사실상 표준으로 자리잡게 되었다. 하둡은 GFS의 구조를 본떠 설계된 분산 파일 시스템인 HDFS(Hadoop Distributed File System)으로 여러 대의 데이터노드(DataNode)에 파일을 저장하며, 파일의 위치 및 권한, 사용자와 같은 메타정보를 네임노드(NameNode)에서 관리하는 구조로 이루어져 있다. 또한 하둡은 분산 처리 프레임워크 Hadoop MapReduce를 포함하고 있어, 대규모의 데이터에 분석 알고리즘을 효과적으로 분산 수행할 수 있다. HDFS와 Hadoop MapReduce라는 두 가지 중요한 구성요소를 통해, 하둡 시스템은 대용량 데이터의 저장 및 분석을 단일 시스템의 성능을 늘리는 스케일-업 방식으로 해결하는 것이 아니라 일반적인 성능을 가진 머신을 다수 연결하여 파일 저장 및 분석을 수행하는 스케일-아웃

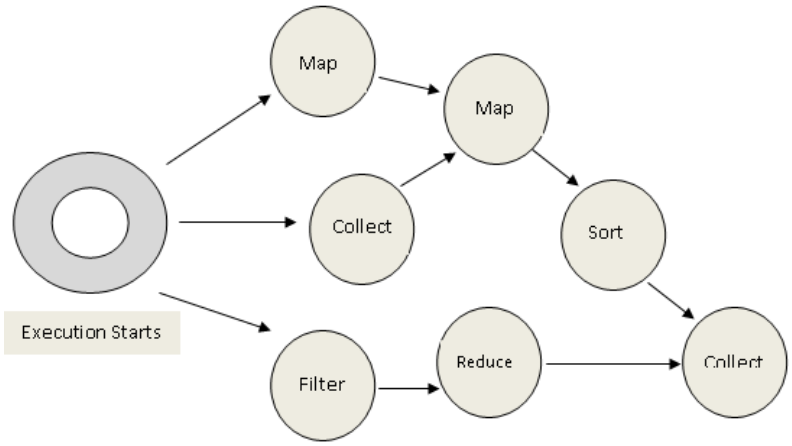
(scale-out) 방식으로 규모를 확장한다.



[그림 3] – 하둡 프레임워크 구조도

분산 처리 프레임워크 영역에서 하둡 생태계의 표준화가 이루어지면서, 하둡의 분산 프로그래밍 모델인 Hadoop MapReduce에 대한 성능 문제가 제기되었다. Hadoop MapReduce의 수행 방식은 우선 클러스터에서 데이터를 읽어 들여 한 단계의 동작을 수행하고 결과를 다시 클러스터에 저장한 다음, 다음 단계의 동작을 수행하기 위해 다시 데이터를 읽어 들이고 결과를 클러스터에 저장하는 방식을 반복한다. 따라서 데이터 처리 또는 분석 알고리즘의 매 단계마다 데이터를 디스크에 쓰고 읽는 과정을 거치기 때문에 디스크 I/O(Input/Output) 수행 시간이 전체 알고리즘 수행 시간에서 많은 부분을 차지하게 된다.

이러한 문제를 해결하기 위해 불필요한 디스크 접근을 최대한 배제하고 메모리에 적재된 데이터를 업데이트하는 과정을 메모리 내에서 수



[그림 4] – 스파크 DAG 예시

행하기 위해 다양한 솔루션이 등장하였으며, 그 중 아파치 스파크(Apache Spark) [41] 프레임워크가 사실상 표준화되었다. 아파치 스파크는 데이터를 메모리에 적재하고 맵/리듀스 등의 분산처리를 메모리 상에서 지원하기 위해 RDD(Resilient Distributed Datasets)라는 데이터 구조를 사용하며, 분산 처리하기 위한 프로그램의 각 단계별 연산을 DAG(directed acyclic graph)로 묘사하고 최적화한다. 이러한 구조를 통해 스파크에서는 한 단계의 연산이 끝난 결과를 디스크에 바로 저장하는 것이 아니라 우선 데이터를 적재한 초기 단계의 RDD부터 모든 연산이 완료된 상태의 RDD까지의 계보(lineage)를 작성하고, 분산 저장된 데이터를 RDD의 계보를 추적하며 메모리 상에서 연산이 수행되도록 한다. DAG와 RDD라는 요소로 인해, 스파크에서는 연산 중간에 분산 처리가 실패할 경우 해당 RDD 조각을 재구성하기 위해 다시 계보를 추적하며 디스크 읽기부터 인메모리 연산을 다시 시작할 수 있어 결함 감내

성(fault tolerance)이라는 특징 또한 가지고 있다.

2.2. K-Means 클러스터링

K-Means 클러스터링은 기계 학습 알고리즘 중 자율 학습(unsupervised learning)의 일종으로, 각 데이터 포인트에 대해 주어진 레이블(label)이 없이 주어진 데이터를 가장 잘 설명하는 클러스터를 찾는 문제이다. 데이터를 분류하기 위해서는 각 데이터에 대해 레이블이 필요하지만, 일반적으로는 데이터의 레이블이나 카테고리를 알지 못하는 경우가 많으므로 분류가 아닌 다른 방식을 통해 데이터를 설명해야 하는 경우가 생긴다.

클러스터를 정의하는 방법 중 가장 직관적인 것은 같은 클러스터에 속한 데이터들이 서로 가깝다고 정의하는 것이다. 즉, 내부 거리가 가장 가까워지도록 클러스터를 고르는 것이 된다. K-Means 클러스터링은 각 클러스터에 대해 중심점을 두고, 클러스터 내의 데이터들이 중심점과 얼마나 가까운지를 목적 함수로 하여 이 값을 줄이는 알고리즘이다. 알고리즘 목적 함수 수식은 [수식 1]과 같다.

$$\min_{b,w} \sum_{i}^{n} \sum_{j}^{k} w_{ij} \|x_i - b_j\|_2^2 \text{ s.t. } \sum_{j} w_{ij} = 1, \forall j$$

[수식 1] - K-Means 클러스터링 목적 함수

K-Means 클러스터링 알고리즘에서 데이터 간의 가까운 정도는 유클리디언 거리로 측정한다. 데이터가 n개, 클러스터가 k개 있다고 하면 위와 같이 클러스터 중심 b_j와 점 x_i가 가까운 정도를 모두 측정하여 더하게 된다. 이 때 w_{ij}는 데이터 i가 클러스터 j에 속하는지 여부를 1 또는 0으로 나타내는 이진 변수이다.

위 목적 함수는 이진 변수 w_{ij} 때문에 최적화하는 것이 어렵다. 각 데이터 x_i에 대한 클러스터 조합을 모두 검증해야 하므로, 목적 함수를 곧바로 최적화하는 것은 계산 불가능하다. 그러나 w_{ij}와 b_j 중 하나를 고정하고 다른 하나를 최적화하는 방식으로 번갈아 최적화를 수행하면 계산이 가능해지는데, 이 때 w_{ij}를 고정하고 b_j를 최적화하기 위해서는 j 클러스터의 모든 점을 평균하여 b_j를 결정한다. 따라서 총 k개의 평균을 클러스터의 중심으로 삼는다. 반대로 b_j가 고정되면, 모든 데이터 점과 각 클러스터 중심 사이의 거리를 다시 측정하여 가장 가까운 중심점을 가진 클러스터로 데이터를 재배치하여 w_{ij}를 갱신할 수 있다. 이 과정을 반복하여 일정 횟수 이상 반복하거나 목적 함수의 값이 거의 변하지 않는 등의 정지 조건을 만족시키면 클러스터를 확정하게 된다.

제 3 장 연구 문제

제 1 절 연구 문제

이 절에서는 대규모 TV 시청 로그를 통해 분석하고자 하는 행동 기반 시청 패턴에 대한 연구 문제를 설정하고, 각 연구 문제에 대해 설명함과 동시에 연구문제 간의 관계에 대해 설명하고자 한다. 선행 연구의 이론적 배경에서 논의했듯이, 채널 선택에 대한 능동성과 콘텐츠 소비에 대한 다양성이 다채널 환경에서의 TV 시청 행위를 이해하는 데에 있어 중요한 틀임을 확인할 수 있었다. 또한 기존의 웹 사용 마이닝 분야에서 사용되어온 세션 기반 클러스터링 기법이 TV 시청 로그에 대해서도 유효한지에 대한 가능성을 확인해보고자 한다.

[연구 문제 1] 세션 기반 클러스터링 기법을 통해 TV 시청 로그에서 시청 행동을 유형화할 수 있는가?

1.1. 유형화 된 시청 행동은 집중적으로 이용하는 채널의 종류에서 어떠한 차이를 보이는가?

1.2. 유형화 된 시청 행동은 콘텐츠 소비에 대한 다양성에서 어떠한 차이를 보이는가?

연구 문제 1에서는 우선 제안된 세션 기반 클러스터링 기법이 TV

시청 로그에 대해서도 유효한 결과를 도출하는지에 대해 알아보고자 한다. 이를 확인하기 위해, 제안된 방식으로 도출된 각 세션 클러스터를 유형화했을 때 각 세션 유형이 기존의 다채널 환경에서 중요한 의미를 가지는 콘텐츠 소비에서의 채널 종류와 다양성 측면에서 TV 환경의 시청 행위에 대한 이해가 가능하도록 구분되는지에 대해 알아보고자 한다.

특히 본 연구에서는 시청 행동의 유형을 통해 구성된 시청 가구의 TV 시청 행태에서 각 가구 유형별로 능동성의 차이를 보이는지에 대해 파악해보고자 하는데, 각 시청 가구가 아닌 시청 행동을 유형화하는 단계에서는 행위 자체의 의도를 명확히 할 수 없다. 따라서 이 단계에서는 행동의 능동성을 논하기보다 일차적으로 케이블 및 지상파, 종합편성채널 등의 채널 종류 중 어떤 카테고리를 집중적으로 이용하는지에 대해 알아보고자 한다.

[연구 문제 2] 유형화 된 시청 행동의 구성비로 TV 시청 가구를 묘사하여 시청 가구를 유형화할 수 있는가?

2.1. 유형화 된 시청 가구는 콘텐츠 소비에 대한 능동성에서 어떠한 차이를 보이는가?

2.2. 유형화 된 시청 가구는 콘텐츠 소비에 대한 다양성에서 어떠한 차이를 보이는가?

연구 문제 2에서는 연구 문제 1을 통해 도출된 시청 행동 유형이

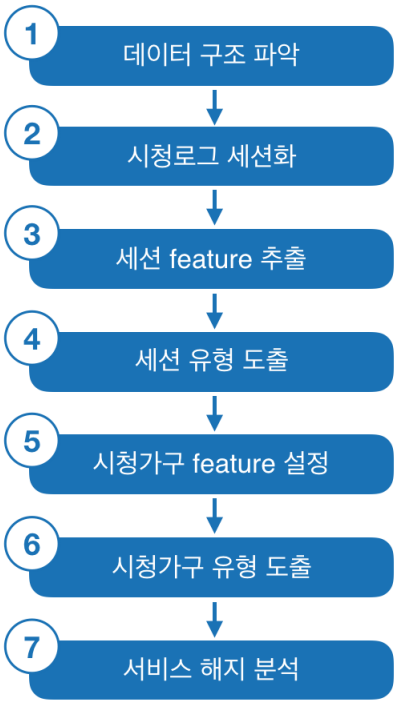
의미 있는 시청 가구 유형을 도출할 수 있는지에 대해 알아보고자 한다. 이를 위해 각 시청 가구가 시청 로그를 통해 보인 시청 행동들을 각 시청 행동 유형에 대한 구성비로 묘사하고, 이를 클러스터링해 시청 가구 유형을 도출해본다. 각 시청 가구 유형이 콘텐츠 소비에 대한 능동성 및 다양성 측면에서 어떠한 의미를 가지는지에 대해 알아보고자 한다.

[연구 문제 3] 유형화 된 시청 행동 및 시청 가구가 서비스 해지와 연관이 있는가?

연구 문제 3에서는 앞서 도출된 시청 행동 및 시청 가구 유형이 TV 시청 행위에 대한 실증적 이해와 더불어 산업적 시사점을 도출할 수 있는지에 대해 알아보고자 한다. 이에 따라, 디지털 케이블 TV의 서비스 해지 기록을 활용하여 각 시청 가구 유형별 서비스 해지율과 가구 유형별 시청 행동 유형 구성비 간의 상관관계를 파악하여, 본 연구에서 제안한 분석 프레임워크가 산업적으로 유효한 결과를 도출 가능한지에 대해 알아보고자 한다.

제 4 장 연구 방법

본 연구는 [그림 5]와 같은 순서로 진행되었다. 먼저 수집된 데이터를 전처리 하여 시청 로그의 형태를 세션 구성에 용이하도록 변경한다. 그 후 전처리 된 시청 로그를 사용하여 각 시청 가구별로 TV 시청에 대한 세션을 구성한 다음, 세션의 특성을 잘 나타낼 수 있는 특질을 추출하여 이를 바탕으로 세션을 클러스터링하고 분석 작업을 통해 세션을 유형화한다. 그리고 유형화 된 세션으로 시청 가구를 모델링하여 클러스터링을 진행한 후 분석하여 시청 가구를 유형화한다.

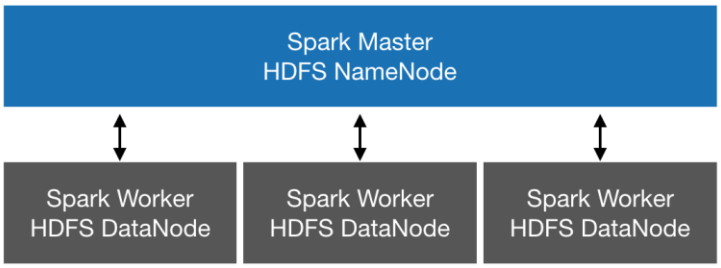


[그림 5] - 연구 방법 개괄

제 1 절 데이터 개요 및 전처리

1.1. 실험 환경

본 연구에 사용된 데이터는 디지털 케이블 TV 서비스인 CJ 헬로tv의 모 지역 가입 가구 시청 로그다. 전체 26만 4,050가구에 대해 2016년 특정 달의 시청 로그 및 다음 달의 서비스 해지 기록을 기반으로 연구를 진행했으며, 총 193,218,856개 시청 로그가 전처리에 사용되었다. 각 가구 별 평균 시청 로그 수는 339.29개이며, 표준편차는 287.35개이다. 연구에 사용된 시청 로그의 크기가 방대하므로, 본 연구를 원활히 진행하기 위해 빅데이터 처리 프레임워크를 구성하였다. 총 4대의 서버 컴퓨터로 구성된 빅데이터 처리 프레임워크는 분산 환경에서 시청 로그 파일을 저장하기 위해 Hadoop 2.7 버전의 HDFS(Hadoop Distributed File System)을 사용하였으며, 대규모 데이터 전처리 및 클러스터링 등을 수행하기 위해 분산 컴퓨팅 플랫폼인 Apache Spark 2.0.2 버전을 사용했다. 대략적인 프레임워크의 구조도는 [그림 6]와 같다.



[그림 6] - 빅데이터 처리 프레임워크 구조도

1.2. 데이터 개요

시청 로그 원 데이터는 [표 1]과 같은 형식으로 구성돼 있다. 시청 가구를 특정하기 위한 암호화된 셋톱박스 하드웨어 주소와 시청한 채널 번호, 그리고 한 시간 단위의 각 시간대별로 해당 채널을 시청한 시각이 분 단위로 기록된다.

	셋톱박스 하드웨어 주소	채널	시청 시간대	분 단위 기록
0	7918fe23ec576cae	153	2016-04-15 22:00:00	0,1,2,3,4,5,6,7,8,9,10
1	7918fe23ec576cae	839	2016-04-15 23:00:00	45,46,47,48,49,50,51

[표 1] – 시청 로그 원 데이터 예시

CJ 헬로tv의 셋톱박스는 한 시간동안 채널 변경 및 머무른 시각을 집계해 한 시간마다 서버로 로그를 전송한다. 따라서 여러 시간대에 걸쳐 연속적으로 같은 채널을 시청한 경우 [표 2]와 같이 각 시간대 단위로 로그가 끊어져 쌓이게 된다. 본 연구에서는 여러 시간대에 걸친 채널 시청이 발생한 경우 데이터를 처리하기 용이하도록 채널 시청 로그를 전처리 했다.

	셋톱박스 하드웨어 주소	채널	시청 시간대	분 단위 기록
0	7918fe23ec576cae	153	2016-04-15 22:00:00	15,16,17,18,19,…,58,59
1	7918fe23ec576cae	153	2016-04-15 23:00:00	0,1,2,3,4,5,…,57,58,59
2	7918fe23ec576cae	153	2016-04-16 00:00:00	0,1,2,3,4,5,6,7,8,9

[표 2] - 시간대 단위로 끊어진 시청 로그 예시

1.3. 데이터 전처리

원시 시청 로그 형식은 곧바로 연구에 적용하는 데에 어려움이 있으므로, 본 연구에서는 우선 시청 로그 형식을 다루기 쉽도록 전처리 하였다. 전처리는 [표 1]의 원시 데이터 형식을 〈가입가구ID, 시청한 채널, 시청 시작 시간, 시청한 시간〉과 같은 형식으로 변환하도록 수행했다.

[그림 7]은 원시 데이터를 전처리한 결과와 비교한 것이다. 여러 시간대에 같은 채널의 로그가 걸쳐 있는 경우에도 시청 로그의 정보를 통합하여, 전처리 로그가 시청 행위를 분절되지 않게 표현할 수 있도록 하였다.

셋톱박스 하드웨어 주소	채널	시청 시간대	분 단위 기록
7918fe23ec576cae	153	2016-05-15 22:00:00	0,1,2,3,4,5,⋯,54,55,56
7918fe23ec576cae	856	2016-05-15 22:00:00	57,58,59
7918fe23ec576cae	856	2016-05-15 23:00:00	0,1,2,3,4,5,⋯,23,24,25
7918fe23ec576cae	894	2016-05-16 08:00:00	14,15,16,17,18,19,20

셋톱박스 하드웨어 주소	채널	시청 시작 시간	시청한 시간(분)
7918fe23ec576cae	153	2016-05-15 22:00:00	57
7918fe23ec576cae	856	2016-05-15 22:57:00	29
7918fe23ec576cae	856	2016-05-16 08:14:00	7

[그림 7] - 데이터 전처리 과정

제 2 절 세션 유형화

이 절에서는 앞서 전처리 된 시청 로그를 이용하여 시청 행위를 클러스터링하는 과정에 대해 살펴본다. 전처리 된 시청 로그는 우선 시청 흐름이 시작될 때부터 끝날 때 까지를 단위로 하는 세션(session)으로 통합된다. 그리고 각 세션에서 시청 행위를 잘 묘사할 수 있는 특질을 추출하여 세션 벡터(vector)를 생성한 후, 모든 세션 벡터에 대해 K-Means 클러스터링을 수행하여 최적의 클러스터를 도출했다.

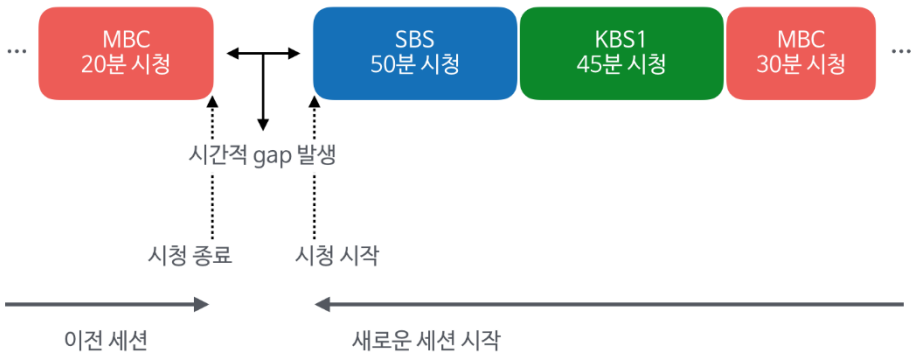
2.1. 세션화

세션(session)은 운영체제나 서버 등 컴퓨터 시스템 관리자의 자산을 사용자가 이용하는 데에 있어, 정보를 상호 교류하는 반영구적인 상태를 의미하는 용어이다. 세션은 사용자가 관리자에게 보낸 연결 요청이 수락된 때에 성립하여 일정 시간 후에 종료된다. 본 연구에서는 디지털 케이블 TV 환경에서 TV 시청 행위를 모델링하기 위해 이 세션 개념을 차용했다. 각 시청 가구가 발생시킨 TV 시청 로그를 통합하여 복수 개의 세션으로 변환하였는데, 이 때 세션의 성립과 종료 시점에 대해 기준을 두어 시청 행위를 세션으로 묘사하고자 하였다.

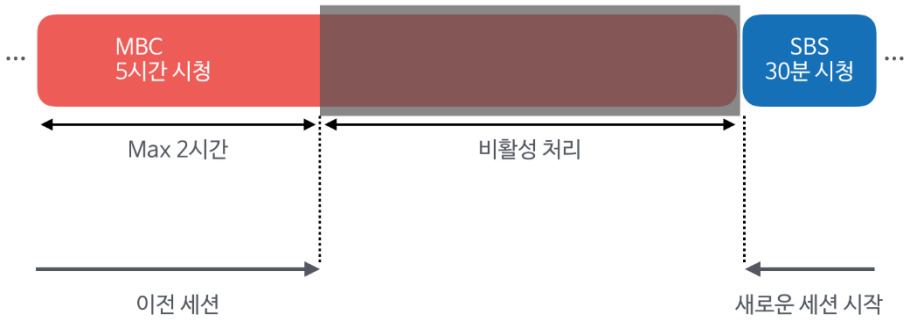
기본적으로 TV 시청 행위에 대한 세션은 시청 가구가 TV 및 셋톱박스의 전원을 켜 TV 시청을 시작할 때 성립하며, TV 및 셋톱박스의 전

원을 꺼 TV 시청을 그만두는 시점에 종료된다. 이 경우 세션 구성은 [그림 8]과 같이 이루어진다. 하나의 세션에서는 다수의 시청 행위가 발생할 수 있는데, 시청할 프로그램을 고르기 위해 채널을 탐색하는 과정과, 한 채널에 안착하여 프로그램을 시청하는 행위가 복수 개 포함될 수 있다.

그러나 디지털 케이블 TV 환경의 경우 셋톱박스의 전원을 그대로 켜 둔 상태에서 셋톱박스와 연결된 TV의 전원을 끄는 것만으로도 시청 행위를 종결할 수 있다. 이 경우 TV의 종료 신호가 셋톱박스에 전달되지 않으므로, 케이블 TV 서비스 공급자의 서버에서는 셋톱박스가 여전히 콘텐츠 서버와 연결돼 있는 것으로 간주하고 지속적으로 콘텐츠를 전송하고 셋톱박스의 시청 로그를 받아들이게 된다. 즉, 서버의 입장에서는 여전히 해당 시청 가구가 시청을 지속하고 있는 것으로 간주된다.

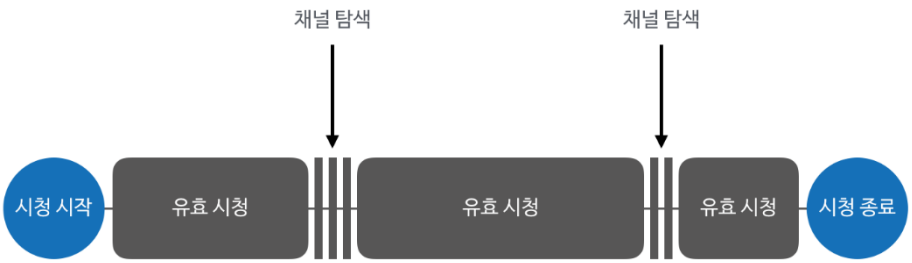


[그림 8] - 셋톱박스의 종료/시작에 따른 세션 구분



[그림 9] – 비활성 시청 행위에 따른 세션 구분

따라서 이러한 경우, 일반적으로 인터넷 이용에 있어 오랜 시간 행동이 비활성화된 경우 세션을 종료하는 것과 마찬가지로 본 연구에서도 일정 시간 이상 같은 채널에 오래 머무르는 행위가 관측될 경우 시청 행동이 비활성화된 것으로 간주하여 세션을 종료하였다. 이 때의 세션 구성은 [그림 9]와 같은데, 일정 시간 이상 비활성화된 경우 세션을 종료하고, 이후 셋톱박스 로그 상으로 채널이 변경된 경우 다시 시청 행위가 활성화된 것으로 보아 새로운 세션을 성립하도록 하여 세션을 구성하였다.



[그림 10] – TV 시청 세션의 구성요소

위와 같은 세션화에 의해, 각 시청 가구의 시청 로그를 연속적인 채널 변환에 대해 통합하여 시청 시작부터 종료를 단위로 하는 세션을 구성하였다. 세션의 각 구성요소는 [그림 10]에서 볼 수 있는데, 시청 시작과 종료 사이에 일어난 채널 변환 및 시청을 채널 탐색과 유효 시청으로 구분하였다. 채널 변환 후 5분 이내에 다른 채널로 넘어가게 되는 경우를 채널 탐색 단계로 정하였다. 그리고 5분 이상 한 채널에 머물러 시청한 경우를 유효 시청으로 보아, 채널 당 시청 시간 등을 계산할 때 채널 탐색 단계를 제외하고 유효 시청인 경우만으로 한정하였다.

2.2. 세션 특질 추출

시청 로그를 세션으로 변환한 후, 각 세션의 특징적인 값을 추출하여 세션 특질 벡터(session feature vector)를 구성하였다. 세션의 특질은 각 세션들을 대표할 수 있고 서로 다른 특성을 가진 세션들이 구분될 수 있도록 선정하였다. 본 연구에서 사용한 세션의 특질은 다음과 같다.

- 세션의 길이
- 총 채널 유효 시청 수
- 채널 유효 시청 시간 총합
- 채널 유효 시청의 평균 시간
- 한 채널 당 평균 시청 시간
- 채널 집중도

- 채널 당 유효 시청이 일어난 수
- 시청한 채널 개수
- 한 프로그램 당 평균 시청 시간
- 한 채널 당 시청한 평균 프로그램 수
- 세션에서 시청한 지상파/종합편성채널/케이블 채널 수
- 세션에서 시청한 지상파/종합편성채널/케이블 프로그램 수
- 총 시청시간 중 지상파/종합편성채널/케이블 채널의 비율
- 케이블 채널의 장르 집중도

여기서 채널 유효 시청이란 시청자가 채널을 탐색하다가 특정 채널에 멈춰 5분 이상 시청한 경우를 말한다. 본 연구에서 사용한 시청 로그는 측정의 최소 단위가 1분이므로 수 초 안에 일어나는 채널 재핑(zapping) 현상은 파악할 수 없다. 그러나 일반적으로 한 프로그램의 방영 시간이 전후 광고 송출 시간을 포함하여 한 시간 내외인 것을 감안하면, 5분 미만으로 채널을 시청하고 다시 채널을 돌리거나 TV 시청을 종료하는 것은 채널의 탐색 과정으로 볼 수 있다. 따라서 본 연구에서는 채널의 유효 시청에 한해서 채널 및 프로그램 시청 시간을 측정하였다.

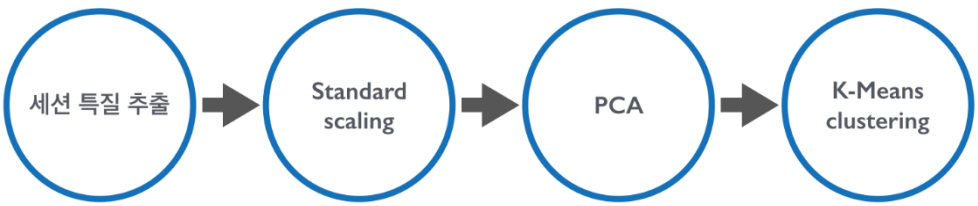
또한 위의 특질에서 채널 집중도란, 시청 시간이 가장 긴 채널의 전체 시청 시간 대비 비율을 백분위로 나타낸 것이다. 이 값은 총 시청 시간이 고정되어 있을 때 가장 오래 시청한 채널의 시간이 길수록 커지므로, 해당 세션의 시청자가 특정 채널에 오래 머물렀는지 혹은 여러 채널

을 두루 보았는지에 대한 값으로 볼 수 있다.

마찬가지로 케이블 채널의 장르 집중도란 케이블 채널의 전체 시청 시간에 대해 가장 오래 시청한 채널 장르의 시간적 비율을 뜻한다. 여기서 채널의 장르란 일반적으로 케이블 TV에서 구분하는 드라마 채널/영화 채널/엔터테인먼트 채널/레저 채널/종교 채널 등을 뜻한다. 이 값을 통해 해당 세션이 특정 장르에 집중하여 케이블 채널을 시청했는지 혹은 다양한 장르를 시청했는지 알 수 있다.

2.3. 세션 클러스터링

앞 단계에서 각 세션을 세션 특질 벡터로 변환한 후, 전체 특질 벡터에 대해 K-Means 클러스터링을 수행하여 클러스터를 도출하였다. 클러스터링을 수행하기에 앞서, 먼저 클러스터링의 품질을 높일 수 있도록 벡터 스케일링 및 PCA(Principal Component Analysis)를 수행하였다. 전체 과정은 [그림 11]과 같다.



[그림 11] - 세션 클러스터링의 전체 과정

우선 특질 벡터에 스케일링을 수행하였다. K-Means 클러스터링의 품질을 높이기 위해서는 각 특질을 동일한 단위로 맞춰 주어야 한다. 본 연구에서 설정한 특질은 각기 단위가 다르므로, 각 차원이 비슷한 통계적 특성을 가지도록 하기 위해 평균을 0, 표준편차를 1로 가지도록 스케일링 하였다. 스케일링에 사용한 식은 [수식 2]와 같다.

$$x' = \frac{x - \bar{x}}{\sigma}$$

[수식 2] - 특질 스케일링 수식

다음 단계로 스케일 된 특질 벡터에 PCA를 수행하였다. 특질 벡터의 차원 수가 20이고, 차원 수가 클 경우 각 데이터 포인트가 데이터 공간의 경계로 향하여 유클리드 거리(Euclidean distance)를 사용해 데이터 포인트의 가까움을 측정하기 어려워지는 차원의 저주(curse of dimensionality) 현상이 발생할 수 있다. 따라서 본 연구에서는 데이터의 차원 축소에 널리 사용되는 PCA 기법을 적용하였다. PCA는 각 특질 축의 선형 조합을 통해 데이터의 분산이 가장 크게 나타나는 축을 찾고, 이와 수직이며 분산이 가장 크게 나타나는 다음 축을 찾는 과정을 반복하여 차원이 축소되더라도 데이터 분포 특성의 유실이 적게 일어나도록 하는 차원 축소 기법이다. 여기에서는 PCA를 통해 20차원의 특질 벡터를 4차원으로 줄여 사용하였다.

마지막으로 차원 축소된 벡터에 K-Means 클러스터링을 적용하여

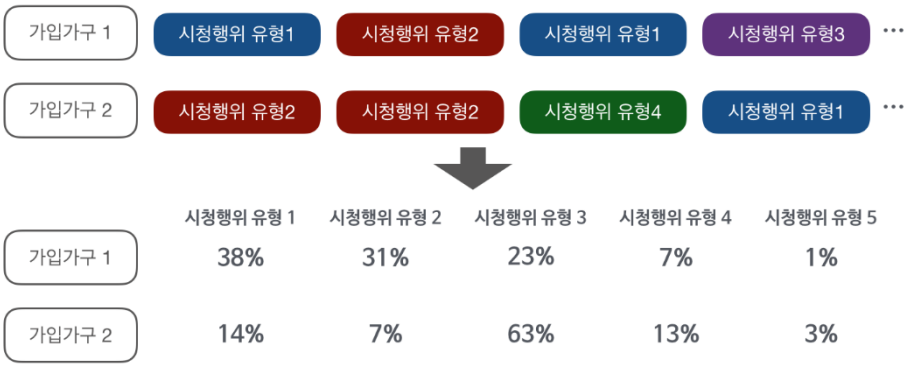
클러스터를 도출하였다. K-Means 클러스터링 기법은 클러스터의 수 K를 지정해야 하는데, 이 때 일반적으로 사용되는 방법은 클러스터 내 점 사이의 거리 합(WSSSE; Within-Subject Sum of Squared Error)이 클러스터 수를 늘려 감에 따라 감소하는 추이를 보고 감소하는 경향이 줄어드는 지점의 클러스터 수를 채택하는 elbow method이다. 본 연구에서도 elbow method를 적용하여, 클러스터의 수에 따라 WSSSE의 값을 scree plot으로 나타내어 WSSSE의 감소가 둔화되는 지점을 최적의 클러스터 수로 선택하였다.

제 3 절 시청 가구 유형화

앞 절에서 유형화한 세션 클러스터를 바탕으로 시청 가구를 모델링하였다. 시청 가구가 발생시킨 각 세션에 앞서 도출한 세션 유형 정보를 입혀, 시청 가구가 각 세션 유형을 어떤 비율로 발생시켰는지를 시청 가구의 특질로 삼았다. 이 특질에 기반한 시청 가구 벡터 집합에 K-Means 클러스터링을 수행하여 최적의 시청 가구 클러스터를 도출했다.

3.1. 시청 가구 특질 추출

시청 가구를 클러스터링하고 유형화하기 위해 시청 가구의 특성을 잘 나타내는 벡터를 구축할 필요가 있다. 본 연구에서는 제안한 바와 같이 세션 유형의 구성비로 시청 가구 벡터를 표현했다. 대략적인 과정은 [그림 12]에서 볼 수 있다. 먼저 각 시청 가구가 발생시킨 모든 세션을 세션 유형별로 집계한 후, 각 시청 가구의 전체 세션 수 대비 각 세션 유형 수의 비율을 계산한다. 각 유형의 비율 값이 곧 시청 가구를 나타내는 특질로 사용된다. 예를 들어 한 시청 가구가 세션 유형 1, 유형 2, 유형 3을 각각 60%, 30%, 10%씩 발생시켰을 경우, 이 시청 가구의 특질 벡터는 (0.6, 0.3, 0.1)이 된다.



[그림 12] - 세션 유형을 사용한 가입가구의 묘사

3.2. 시청 가구 클러스터링

각 세션 유형의 구성비로 나타낸 시청 가구의 특질 벡터 집합에 대하여 클러스터링 및 유형화를 진행했다. 클러스터링은 2절의 세션 클러스터링과 마찬가지로 K-Means 클러스터링 기법을 사용하였으며, 시청 가구 특질 벡터의 경우 각 특질의 최댓값과 최솟값이 각각 1, 0으로 일정하기 때문에 세션 클러스터링 단계와 같은 특질 스케일링 및 PCA 등의 추가 과정은 수행하지 않았다. 앞과 마찬가지로, K-Means 클러스터링의 클러스터 수는 WSSSE의 scree plot을 그린 후 elbow method를 사용하여 값의 감소가 둔화되는 지점으로 선택하였다.

제 4 절 세션 유형과 서비스 해지의 연관성

앞의 과정을 통해 도출되는 각 세션 유형과 세션 유형의 조합으로 이루어진 시청 가구 유형의 유효성을 평가하기 위해, 본 연구에서는 세션 유형과 서비스 해지 사이의 연관성에 대한 실험을 진행하였다. 서비스 해지 데이터는 [표 3]와 같이, 로그가 수집된 달의 다음 달에 CJ 헬로tv 서비스를 해지한 모 지역 서비스 구독자의 셋톱박스 하드웨어 주소 ID와 서비스 해지일, 그리고 구독을 시작했던 서비스 가입일로 구성돼 있다.

셋톱박스 하드웨어 주소	서비스 가입일	서비스 해지일
723ace92feg154	2013-04-10	2016-04-07

[표 3] – 서비스 해지 데이터 예시

본 연구에서는 세션 유형과 서비스 해지 사이의 연관성을 밝히기 위해, 먼저 각 시청 가구 유형 그룹의 서비스 해지자 비율을 계산한 후, 해지율과 시청 가구 유형의 각 세션 유형 구성비 간의 피어슨 상관계수를 계산하였다. 여기서 문제를 단순화하기 위해, 각 시청 가구 유형을 구성하는 시청 가구들은 2016년의 특정한 한 달 간 보인 세션 유형 구성비를 서비스 시작일부터 가입기간동안 동일하게 유지했다고 가정하였다. 위의 방법으로 도출된 각 세션 유형 구성비와 해지율 간의 상관계수를 통해, 각 세션 유형이 서비스 해지와 어떤 연관성을 가지고 있는지를 파악할 수 있다.

제 5 장 연구 결과

이 장에서는 3장의 연구 방법을 통해 얻어진 세션 및 시청 가구 유형에 대해 살펴본다. 우선 첫 번째 단계로 세션화 된 시청 행위들에는 어떠한 유형이 있는지 살펴보고, 도출된 세션 유형으로 묘사한 시청 가구들이 어떤 시청 패턴을 보이는지에 대해 살펴보도록 한다.

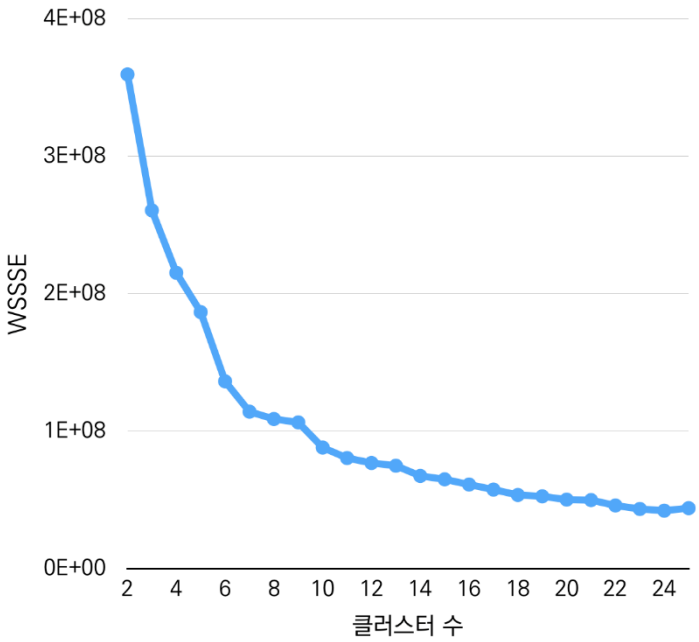
제 1 절 세션 유형

원시 시청 로그에서 TV 시청의 행동적 특성을 유형화하기 위해, 먼저 전체 시청 가구의 시청 로그를 세션화 하였다. 총 193,218,856개의 시청 로그가 TV 시청을 시작한 시점부터 종료한 시점까지의 세션으로 변환되어, 총 29,313,577개의 세션을 구성하고 이를 연구에 활용하였다. 세션 클러스터링에 활용할 각 특질에 대한 전체 세션의 평균 및 표준편차는 다음 표와 같다([표 4]).

특질	평균	표준편차
세션 내 시청 로그 수	2.98	3.73
세션 길이(분)	101.31	56.91
유효 시청 개수	1.91	1.65
평균 유효 시청 시간(분)	69.15	41.73
채널 당 시간(분)	71.29	41.21
채널 집중도(%)	86.44	20.28
시청한 지상파 채널 개수	0.61	0.79
시청한 종편 채널 개수	0.23	0.52
시청한 케이블 채널 개수	0.89	1.00
지상파 프로그램 수	1.14	1.59
종편 프로그램 수	0.34	0.83
케이블 프로그램 수	1.92	2.45
시청시간 중 지상파 비율(%)	37.81	44.85
시청시간 중 종편 비율(%)	11.85	29.26
시청시간 중 케이블 비율(%)	51.34	46.64
케이블 장르 집중도(%)	57.32	46.99
프로그램 당 시청 시간(분)	33.57	19.07
채널 당 시청한 프로그램 개수	2.37	1.92
총 세션 수		**29,313,577**

[표 4] – 전체 세션의 각 특질 별 평균 및 표준편차

전체 시청 가구의 시청 로그를 세션화하고 특질을 추출해 클러스터링하는 과정을 거쳐, 총 7개의 시청 유형을 도출하였다. 세션을 클러스터링하기 위해 사용한 K-Means 클러스터링 기법은 클러스터의 수를 연구자가 정해야 하는데, 본 연구에서는 elbow method를 사용해 WSSSE의 scree plot에서 값이 둔화되는 지점을 최적 클러스터 수로 선정하였다.



[그림 13] - 세션 클러스터링의 WSSSE scree plot

[그림 13]에서, WSSSE 값이 7개의 클러스터부터 둔화되는 것을 볼 수 있다. 따라서 세션 클러스터링 과정에서의 최종 클러스터 수를 7개로 선정하여, 각 클러스터의 특징을 살펴보고 클러스터 중심점(centroid)의 특징을 통해 각 세션 클러스터의 유형을 분류해 보았다. 각 클러스터의 특질 별 평균값은 아래 표에서 볼 수 있다.

세션 유형	케이블 – 배경적	케이블 – 습관적	지상파 – 배경적	지상파 – 습관적	다중– 탐색적	종편– 습관적	다중– 긴 탐색적
구성비	22.47	18.16	16.27	15.63	13.04	10.86	3.57
평균 세션 길이	123.25	48.59	126.77	60.05	111.47	98.54	267.22

평균 세션 로그 개수	1.3	2.29	1.44	2.52	7.21	2.69	11.53
채널 안착 시청 개수	1.12	1.35	1.24	1.69	3.92	1.76	6.93
평균 채널 안착 시청시간	113.69	37.91	109.78	38.6	29.46	68.82	42.41
채널 당 평균 시청시간	114.8	38.57	111.43	39.96	34.38	70.39	55.79
채널 당 안착 시청횟수	1.02	1.03	1.03	1.06	1.22	1.05	1.38
채널 집중도	0.98	0.92	0.96	0.87	0.58	0.87	0.46
시청한 지상파 채널 개수	0.02	0.06	1.13	1.38	0.74	0.21	2.07
시청한 종편 채널 개수	0.01	0.03	0.02	0.08	0.3	1.24	0.81
시청한 케이블 채널 개수	1.07	1.21	0.05	0.12	2.29	0.2	2.25
시청한 지상파 프로그램 개수	0.04	0.06	3.17	1.96	0.97	0.26	4.03
시청한 종편 프로그램 개수	0.01	0.03	0.02	0.08	0.32	2.14	1.18
시청한 케이블 프로그램 개수	4.45	1.88	0.07	0.12	3.06	0.24	3.55
지상파 시청시간 비율	0.01	0.02	0.99	0.96	0.23	0.05	0.49

종편 시청시간 비율	0	0.01	0	0.02	0.07	0.9	0.14
케이블 시청시간 비율	0.99	0.97	0.01	0.02	0.7	0.05	0.36
케이블 장르 집중도	0.99	0.96	0.05	0.12	0.82	0.17	0.67
프로그램 당 평균 시청시간	39.33	26.25	42.88	28.99	24.81	38.15	30.25
채널 당 평균 프로그램 개수	4.26	1.61	2.83	1.41	1.35	1.76	1.82

[표 5] – 세션 유형별 클러스터 중심점 값

다음으로 각 세션 유형에 대해 자세히 짚어보도록 한다. 각 세션 유형의 구성비 및 평균 세션 길이와 함께, 특징적인 시청 행동을 통해 세션 유형을 구분하였다. 특히 지상파와 종합편성채널, 그리고 케이블 채널의 시청량에 따른 차이가 두드러지게 나타났다.

1.1. 케이블-배경적 세션 유형

이 유형은 전체 세션 수의 약 23%를 차지해 가장 큰 구성비를 가진다. 세션 유형의 특징으로는 우선 한 번 시청을 시작하면 약 두 시간 가량 시청을 지속하는 모습을 보인다. 평균 세션 길이는 지상파-배경적, 다중-탐색적 유형과 비슷하나, 이 유형은 주로 케이블 채널 중 하나의 채널에 고정되어 한 채널에서 프로그램을 4-5개 가량 시청하고 시청 활

동을 종료한다는 특징이 있다. 또한 세션이 포함하고 있는 평균 시청 로그 개수가 1.3개인 것으로 미루어 보아, 여러 채널을 탐색하기보다는 곧바로 시청자가 원하는 채널로 이동하여 프로그램 시청을 시작하는 유형의 행동으로 보인다.

1.2. 케이블-습관적 세션 유형

케이블-습관적 세션 유형은 전체 세션 중 18%를 차지하며, 약 50분 가량의 평균 세션 길이를 가진다. 이 유형은 케이블-배경적 유형과 마찬가지로 케이블 채널 영역에 집중된 시청 패턴을 보이지만, 세션 길이가 더 짧음에도 탐색한 채널의 수가 두 배 가량 많은 특징이 있다. 즉, 이 유형은 케이블-배경적 유형과 비교해 보았을 때 채널 탐색에 조금 더 적극적이나, 결국 원하는 프로그램을 1-2개 가량 시청 후 시청을 종료한다.

1.3. 지상파-배경적 세션 유형

전체 세션 중 약 16% 가량을 차지하는 세션 유형으로, 지상파-배경적 세션 유형은 시청을 시작하면 평균적으로 두 시간에 걸쳐 시청을 지속한다. 앞서 살펴보았듯이 이는 케이블-배경적 세션 유형과 비슷하나, 이 유형은 지상파에 집중해 한 채널만을 시청하고 종료한다는 특징이 있다. 이 유형과 케이블-배경적 세션 유형은 단지 주로 시청한 채널의 카테고리만이 다르므로, 만약 채널의 유형을 특질로써 나누지 않았다면,

동일하게 배경적 유형으로 케이블-배경적 유형과 합쳐져서 도출되었을 가능성이 높다.

1.4. 지상파-습관적 세션 유형

다음으로 많은 구성비를 차지한 세션 유형은 지상파-습관적 유형으로, 전체 세션의 약 16%로 구성되었다. 이 유형은 케이블-습관적 세션 유형과 비슷한 특징을 가지나, 두 가지 사뭇 다른 특성을 가지고 있다. 우선 지상파 채널 위주로 시청하는 특성을 가졌으며, 총 시청 시간 대비 가장 오래 시청한 채널의 시청 시간 비율인 채널 집중도가 평균 87%로 케이블-습관적 세션 유형의 92%에 비해 낮은 모습을 보인다. 즉, 케이블-습관적 세션 유형의 경우 한 채널을 줄곧 시청하다 종료하는 반면, 지상파-습관적 유형은 지상파 채널에서 비슷한 행동을 보이지만 비교적 채널 탐색이 더 일어난다고 볼 수 있다.

1.5. 다중-탐색적 세션 유형

다중-탐색적 세션 유형은 전체 세션의 약 13%를 차지한다. 이 세션 유형은 평균 두 시간 정도의 세션 길이를 가지는데, 앞서 살펴본 네 유형과 비교해 봤을 때 매우 적극적으로 채널을 탐색하는 모습을 보인다. 대략 7-8개의 채널을 탐색하며, 그 중 약 4개의 채널에 안착하여 시청을 지속한다. 특징적으로는 전체 시청시간 중 종합편성채널이 차지하는 비율이 극히 적으며, 케이블 채널과 지상파 채널을 7:3 가량의 비율로

시청한다. 또한 채널 집중도가 58%인 점을 미루어 보아, 한 채널에서 오래 머무르지 않고 여러 채널을 균형 있게 시청하는 모습을 보인다.

1.6. 종편-습관적 세션 유형

종편-습관적 세션 유형은 전체 세션의 약 11%를 차지한다. 세션의 평균 길이는 약 한 시간 반 가량으로, 주로 종합편성채널에 집중하여 약 두 개 가량의 프로그램을 시청하는 행태를 보인다. 이 세션 유형의 특성은 케이블-습관적 및 지상파-습관적 유형과도 비슷한데, 다만 특징적인 부분으로 나머지 두 유형이 약 한 시간 가량의 평균 시청 시간을 가지는 것보다는 더 긴 평균 세션 길이를 가진다는 점이 있다.

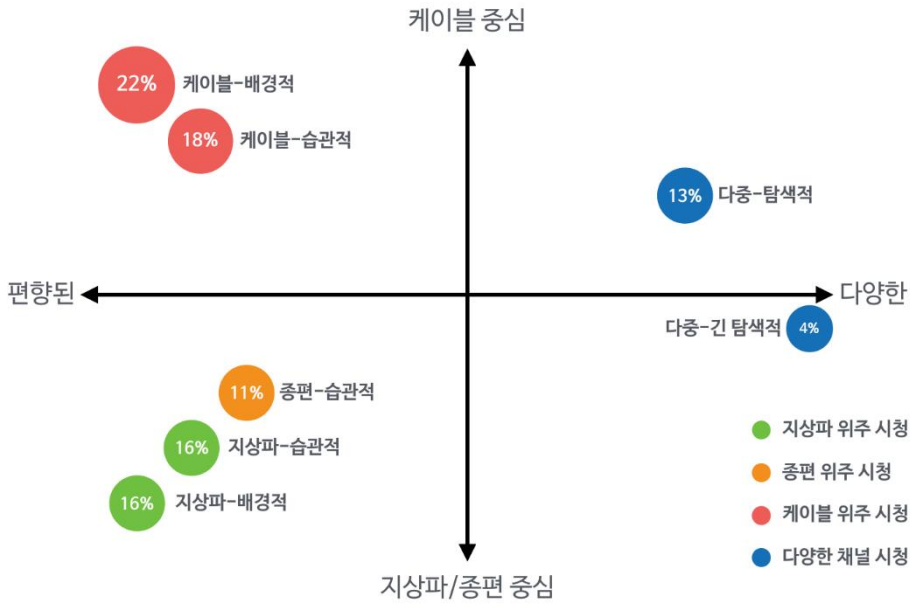
1.7. 다중-긴 탐색적 세션 유형

마지막으로 전체 세션 중 약 4%의 구성비를 가지는 다중-긴 탐색적 세션 유형이 있다. 이 유형은 나머지 유형에 비해 상당히 긴 시청 시간을 보이는데, 평균 세션 시간이 네 시간을 넘기는 것으로 나타났다. 자연스럽게 채널 탐색 횟수도 많아져 평균적으로 약 12개의 채널을 탐색하며, 그 중 7개 가량의 채널을 5분 이상 시청한 것으로 보여진다. 채널 집중도 또한 46%로 한 채널을 오래 보는 편이 아니며, 지상파와 종합편성채널, 그리고 케이블 채널을 각각 5:2:3 정도의 비율로 시청하여 다중-탐색적 세션 유형보다도 더욱 고른 채널 시청 행태를 보였다. 또한 케이블 채널의 드라마, 엔터테인먼트, 레저 등 각 채널 장르 중 가장 오

래 시청한 장르의 전체 케이블 시청시간에 대한 시청 시간 비율인 케이블 장르 집중도 또한 67%로 나타나, 장르적으로도 시청이 편향되지 않고 다양하게 채널을 탐색하는 것으로 보인다.

1.8. 결과 정리

앞서 살펴보았듯, 주로 지상파, 종합편성채널 및 케이블 채널의 집중도에 따른 특성과 함께 세부적으로 총 시청 시간 및 채널 탐색 횟수, 시청한 프로그램 개수 등 다양한 특징에 따라 도합 일곱 개의 세션 클러스터를 유형화 해보았다. 클러스터링 및 유형화 결과에 의하면, 어떠한 채널 카테고리(지상파·종편·케이블)를 주로 시청하는지에 대한 특질이 세션 유형을 나누는 데에 굉장히 중요한 역할을 했다는 것을 알 수 있다. 즉, TV 시청 가구의 시청 행태는 '어떻게 보는가'에 대한 행동적 특성에 앞서, '무엇을 보는가'가 매우 큰 요소를 차지한다고 볼 수 있다.



[그림 14] – 소비 다양성과 주 이용 채널에 따른 세션 유형 분류

일곱 개의 세션 유형을 시청 행위 내에서 보여지는 소비 다양성 및 중심이 되는 채널 카테고리로 구분하여 [그림 14]와 같이 나타냈다. 도출한 클러스터의 각 특질 값에 의하면, 특정 카테고리의 채널을 중점적으로 시청하는 시청 행위는 한 시청이 시작되어 종료될 때까지 다양한 채널과 프로그램을 소비하기 보다는 특정 채널 위주로 머무는 경향이 큰 것으로 나타났다. 이러한 경향은 채널 카테고리의 구분 없이 케이블, 지상파, 종합편성채널 모두에게서 보였다. 특정 카테고리에 대해 크게 집중하지 않고 다양하게 채널을 시청하는 시청 행위 유형(다중-탐색적 및 다중-긴 탐색적 유형)은 세션의 평균 지속 시간에 따라 그 다양성에서 차이를 보인다. 또한 두 유형에게서 시청 시간의 비율이 높은 채널 카테

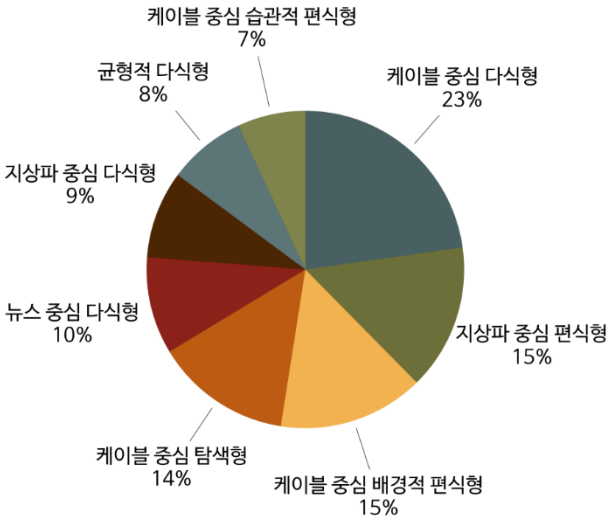
고리가 각각 달랐는데, 먼저 다중-탐색적 유형은 세션 길이가 비교적 짧으며 케이블 채널로 시청 시간 비율이 쏠려 있는 모습을 보였다. 반면 다중-긴 탐색적 유형은 평균 세션 길이가 두 배 이상 길었는데, 앞의 다중-탐색적 유형과 달리 비교적 균형적으로 채널을 이동하는 것으로 나타났다. 또한 케이블 채널만을 대상으로 한 채널 장르의 집중도에서도 다중-탐색적 유형보다 낮은 집중도를 보여, 전반적으로 고른 시청 행위를 보이는 유형으로 해석할 수 있다.

제 2 절 시청 가구 유형

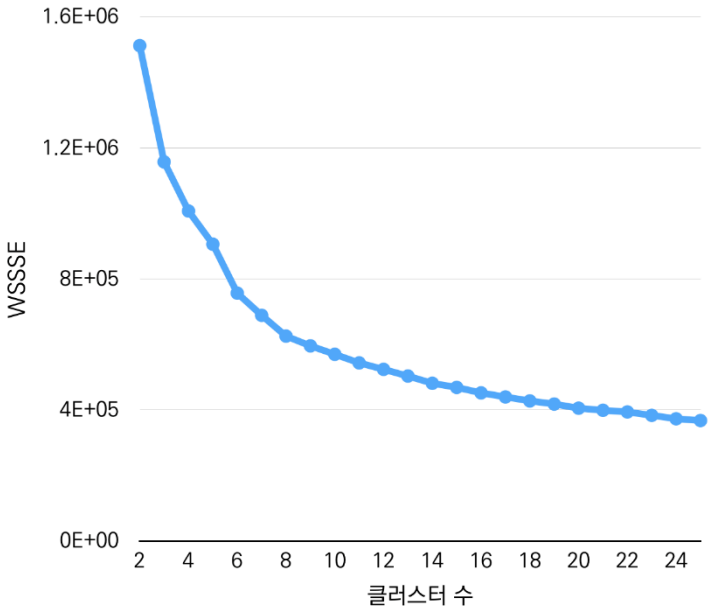
이 절에서는 앞 절에서 도출한 각 세션 유형의 조합 비율로 시청 가구를 묘사하고, K-Means 클러스터링을 사용하여 시청 가구를 세분화한 후 각 시청 가구 클러스터의 시청 패턴에서 나타나는 특징을 바탕으로 시청 가구를 유형화하였다. 클러스터링 결과에 대해 살펴보기에 앞서, 전체 시청 가구가 나타내는 평균 시청 패턴을 [표 6]으로 나타냈다. 각 가구는 평균 약 342개의 시청 로그에 대해 116개의 세션을 생성하였으며, 각 세션은 평균 98분 정도 지속되었다. 전체 가구를 대상으로 시청한 채널의 카테고리를 비교해보면, 케이블 채널의 시청 시간이 가장 많았으며 그 뒤를 이어 지상파 채널, 종합편성채널 순으로 시청 시간 비율이 작아졌다. 또한 데이터 수집 기간인 2016년 특정 한 달 동안 평균적으로 약 35개 정도의 채널을 시청한 것으로 나타났다.

	평균	표준편차
가구 당 로그 수	342.11	293.63
가구 당 세션 수	116.16	54.94
가구 당 평균 세션 지속시간(분)	98.29	50.38
가구 당 지상파 시청 비율(%)	36.19	24.89
가구 당 종편 시청 비율(%)	10.98	12.58
가구 당 케이블 시청 비율(%)	52.83	27.20
가구 당 시청한 채널 수	34.68	21.73

[표 6] – 전체 시청 가구의 시청 패턴 평균 및 표준편차



[그림 15] – 각 시청 가구 유형의 구성비



[그림 16] – 시청 가구 클러스터링의 WSSSE scree plot

세션 클러스터링과 마찬가지로 시청 가구 클러스터링 또한 elbow method를 사용해 최적 클러스터 수를 결정하였는데, 이에 사용된 WSSSE의 scree plot은 [그림 16]와 같다. WSSSE의 scree plot을 살펴보면 클러스터의 수가 8개일 때부터 WSSSE의 값이 감소되는 양이 둔화되기 시작하므로, 본 연구에서는 시청 가구 클러스터의 수를 8로 설정하여 분석을 진행하였다. 클러스터로 도출된 시청 가구 유형의 각 세션 유형 구성비는 [표 7]과 같으며, 이를 구분하기 쉽도록 [그림 17]과 같이 레이더 차트(radar chart)로 나타냈다.

가구 유형	케이블-배경적	케이블-습관적	지상파-배경적	지상파-습관적	종편-습관적	다중-탐색적	다중-긴 탐색적
케이블 중심 다식형	28.29	18.76	17.25	12.97	8.99	11.42	2.32
지상파 중심 편식형	12.02	6.58	45.53	22.64	6.34	4.52	2.36
케이블 중심 배경적 편식형	65.62	22.85	3.89	2.03	2.01	3.19	0.04
케이블 중심 탐색형	14.26	25.78	5.58	13.46	7.88	29.67	3.37
뉴스 중심 다식형	14.00	10.94	13.19	13.13	37.89	8.14	2.70
지상파 중심 다식형	5.31	16.40	11.49	45.04	8.60	10.98	2.18

균형적 다식형	12.75	11.92	14.90	15.63	10.91	16.85	17.04
케이블 중심 습관적 편식형	14.98	51.33	3.11	11.06	4.84	14.00	0.68

[표 7] – 시청 가구 유형의 세션 유형 구성비

도출된 각 시청 가구 유형은 평균 세션 길이 및 지상파·종편·케이블 채널의 시청 시간 비율에서 차이를 보이는데, 이를 [표 8]과 같이 나타냈다.

가구 유형	가구 수 비율(%)	평균 세션 길이(분)	지상파 시청 시간 비율(%)	종편 시청 시간 비율(%)	케이블 시청 시간 비율(%)
케이블 중심 다식형	22.90	102.80	34.74	9.79	55.48
지상파 중심 편식형	15.02	107.78	70.87	7.15	21.98
케이블 중심 배경적 편식형	14.96	105.60	6.85	2.31	90.83
케이블 중심 탐색형	14.19	92.34	26.30	10.12	63.58
뉴스 중심 다식형	9.78	102.03	31.49	37.43	31.08
지상파 중심 다식형	8.63	78.24	61.77	10.93	27.30
균형적 다식형	7.71	132.24	43.90	14.46	41.64

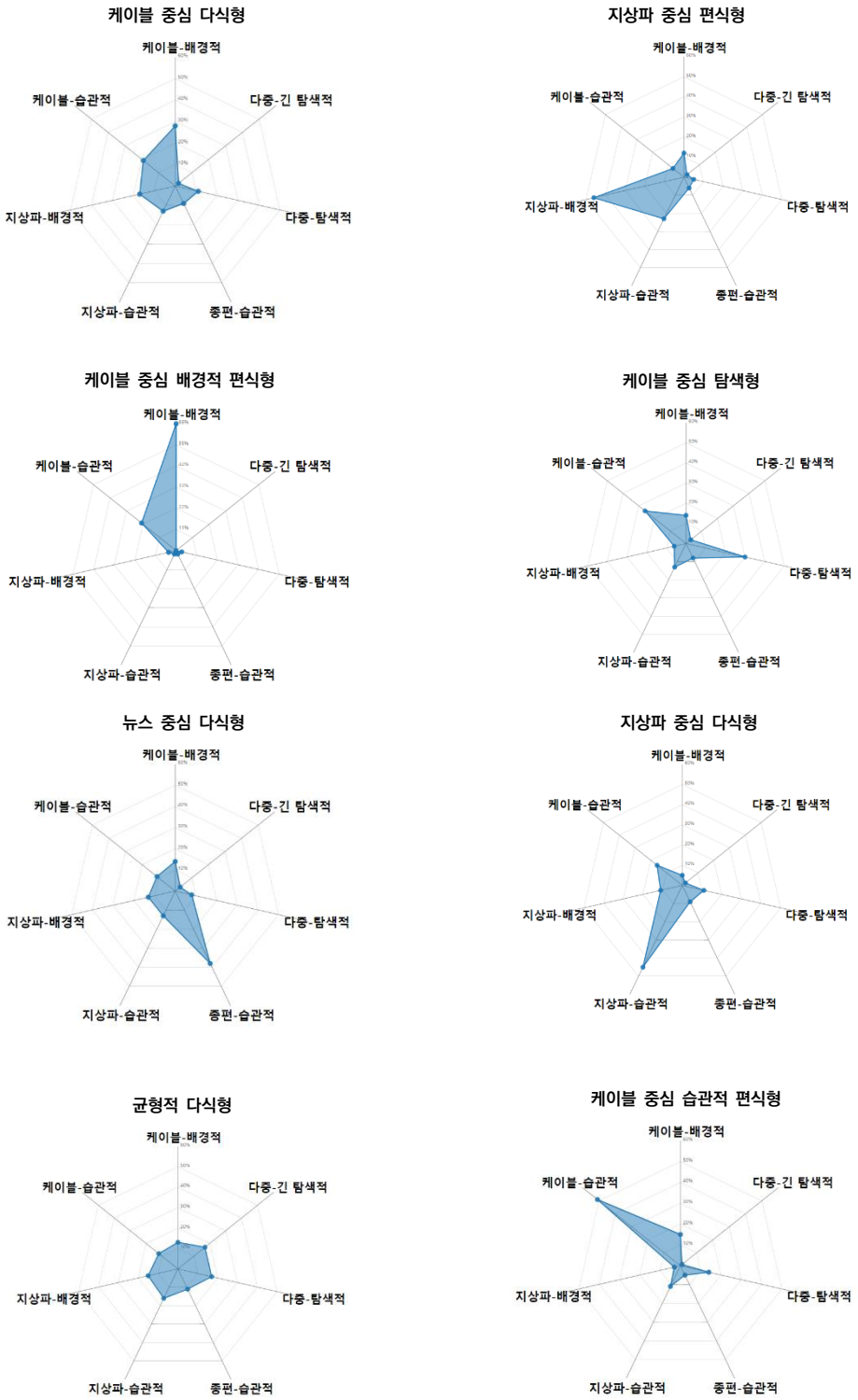
케이블 중심 습관적 편식형	6.80	70.14	17.09	6.08	76.84

[표 8] – 각 시청 가구 유형의 TV 시청 특성

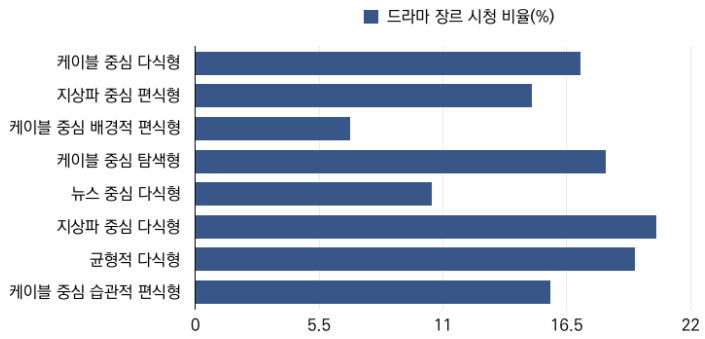
각 시청 가구 유형이 TV 시청 시 나타내는 서비스 이용 특성을 더 구체적으로 알아보기 위해 시청 가구 유형별 케이블 채널의 이용 현황을 도출하여 살펴보았다. 지상파 채널 또는 종합편성채널이 한 채널에서 다양한 장르의 프로그램을 통합적으로 방영하는 것과 달리, 케이블 채널은 각 채널이 특정한 장르(드라마, 엔터테인먼트, 홈쇼핑, 뉴스, 레저 등)의 프로그램을 위주로 방영한다. 본 연구에서는 케이블 채널의 장르 중 상대적으로 시청량이 많으면서 시청 가구 유형별로 시청량의 차이를 보이는 드라마, 뉴스, 홈쇼핑, 엔터테인먼트, 스포츠, 영화, 어린이 채널의 가구 유형별 시청 비율과 함께, 채널의 장르로 구분되지는 않지만 독특한 시청 특성을 보이는 VOD 서비스 광고 채널의 가구 유형별 시청 비율을 함께 고려하여 시청 가구를 유형화하였다. 각 장르 및 VOD 채널의 가구 유형별 시청 비율은 [그림 18]과 같이 막대 그래프로 나타냈다.

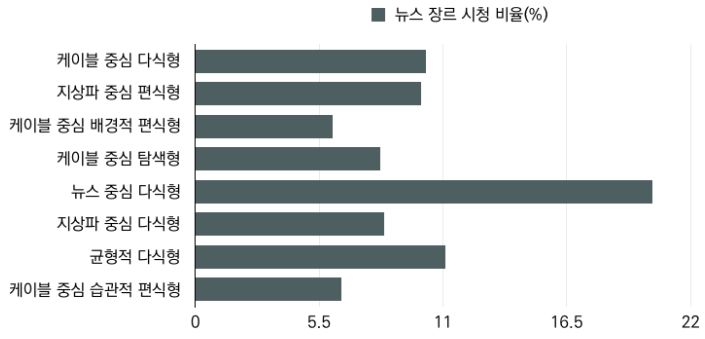
다음은 클러스터로 도출된 각 시청 가구 그룹에 대해 세부적으로 짚어보고자 한다. 각 시청 가구 유형이 주로 어떤 세션 유형으로 조합되어 있으며, 세션 유형의 조합이 지상파 채널과 종합편성채널, 케이블 채널의 시청량으로 드러나는 관계에 대해 살펴본다. 또한 케이블 채널의 각 채널 장르 중 시청량이 높은 드라마, 엔터테인먼트, 뉴스, 스포츠, 홈

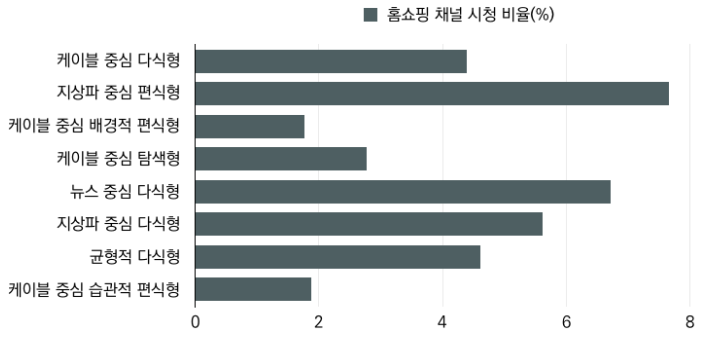
쇼핑, 영화, VOD, 어린이 채널에 대해 각 시청 가구 유형의 시청량을 통한 프로그램 장르 선호도를 알아본다.

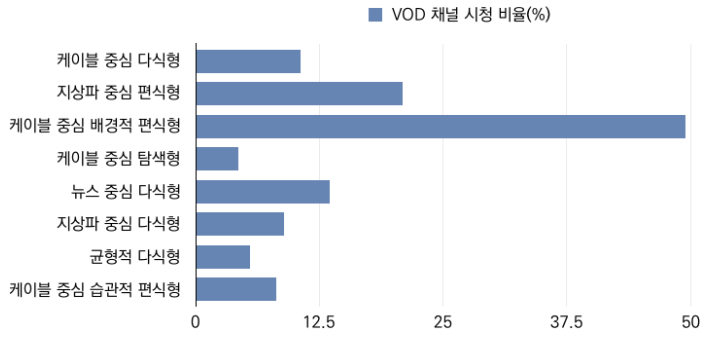


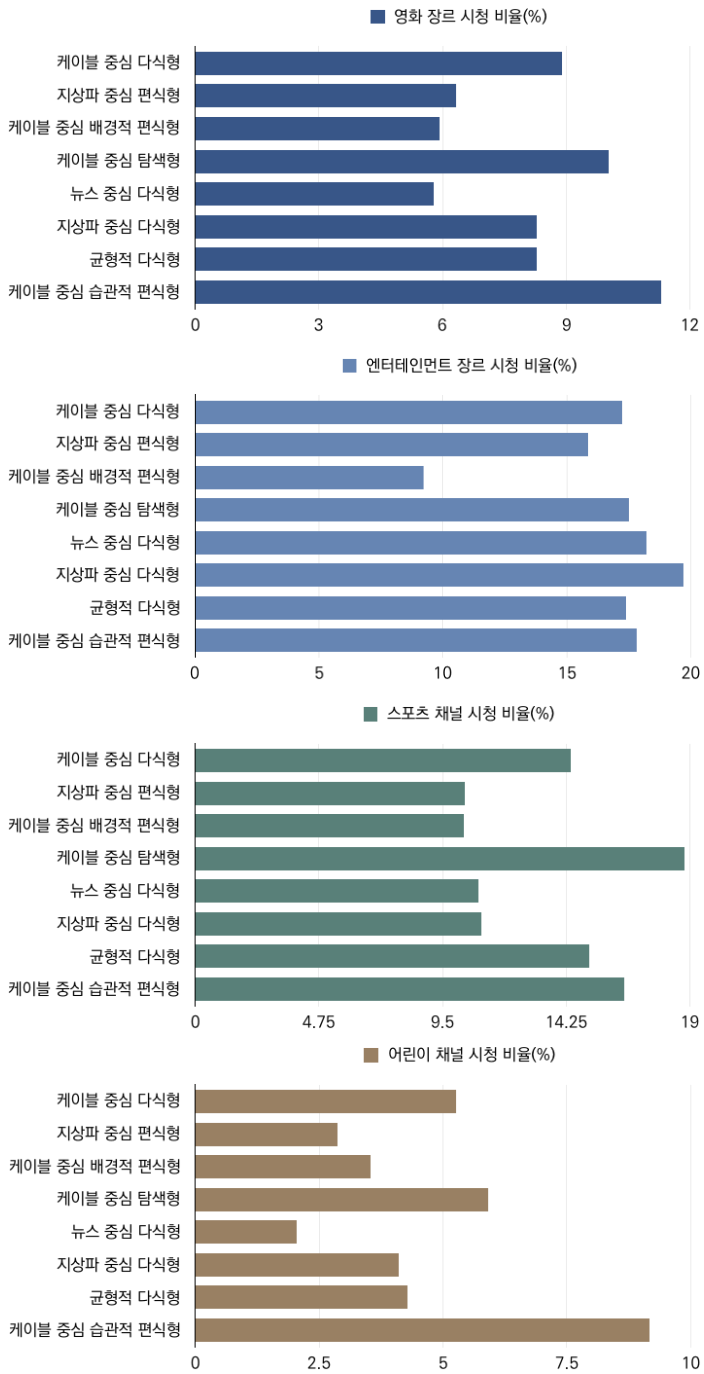
[그림 17] - 각 가구 유형의 세션 유형 조합


드라마 장르 시청 비율(%)
케이블 중심 다식형
지상파 중심 편식형
케이블 중심 배경적 편식형
케이블 중심 탐색형
뉴스 중심 다식형
지상파 중심 다식형
균형적 다식형
케이블 중심 습관적 편식형
0
5.5
11
16.5
22


뉴스 장르 시청 비율(%)
케이블 중심 다식형
지상파 중심 편식형
케이블 중심 배경적 편식형
케이블 중심 탐색형
뉴스 중심 다식형
지상파 중심 다식형
균형적 다식형
케이블 중심 습관적 편식형
0
5.5
11
16.5
22


홈쇼핑 채널 시청 비율(%)
케이블 중심 다식형
지상파 중심 편식형
케이블 중심 배경적 편식형
케이블 중심 탐색형
뉴스 중심 다식형
지상파 중심 다식형
균형적 다식형
케이블 중심 습관적 편식형
0
2
4
6
8


VOD 채널 시청 비율(%)
케이블 중심 다식형
지상파 중심 편식형
케이블 중심 배경적 편식형
케이블 중심 탐색형
뉴스 중심 다식형
지상파 중심 다식형
균형적 다식형
케이블 중심 습관적 편식형
0
12.5
25
37.5
50



[그림 18] – 각 가구 유형별 주요 케이블 장르 시청 비율

2.1. 케이블 중심 다식형

‘케이블 중심 다식형’은 케이블 채널과 관련된 세션의 구성비율이 가장 높고, 그 뒤로 지상파 채널과 종편 채널의 세션 구성비가 따라온다. 여러 카테고리의 채널을 골고루 시청하는 탐색적 세션 유형 중에는 비교적 짧은 탐색적 세션 유형의 구성비가 더욱 높다. 이에 따라 카테고리별 채널 시청량 구성 역시 케이블 채널을 필두로 지상파 채널과 종합편성채널이 뒤따르는 모양새를 보여준다. 케이블의 장르적 선호도를 보면 엔터테인먼트 채널의 선호도가 가장 높으며, 그 뒤를 따라 드라마, 스포츠 및 뉴스 채널의 선호를 보인다. 이는 전형적인 케이블 채널의 분극화 현상을 실증적으로 보인 것이라 할 수 있겠으나, 의외로 VOD 채널의 시청량이 영화 및 홈쇼핑 등의 채널보다 높은 현상 또한 보인다.

2.2. 지상파 중심 편식형

‘지상파 중심 편식형’ 가구 유형은 지상파 채널에 대한 강한 선호를 보이며, 그 외에 종합편성채널보다는 케이블 채널을 더 선호하는 모습을 보인다. 지상파 채널은 주로 배경적 특성을 지닌 시청 행위를 통해 시청하며, 그 뒤를 따라 지상파 채널에서의 습관적 시청 패턴을 보인다. 케이블 채널에서는 앞의 ‘케이블 중심 다식형’ 유형과 마찬가지로 엔터테인먼트와 드라마 장르의 채널을 주로 선호하는데, 마찬가지로 이 유형에서도 VOD 서비스 홍보 채널을 일정량 시청하는 모습을 보여준다.

2.3. 케이블 중심 배경적 편식형

앞서 살펴본 세 유형에 비해 '케이블 중심 배경적 편식형'은 상당히 독특한 세션 유형 조합 및 시청 패턴을 보이는데, 케이블 채널과 관련된 세션 유형에 집중돼 있는 모양새를 보인다. 특히 케이블-배경적 세션 유형의 비율이 매우 높다. 지상파 채널 및 종합편성채널, 그리고 여러 채널 종류에 걸친 세션 유형이 거의 보이지 않는 만큼 채널 시청 경향 또한 케이블 채널로 완전히 집중된 모양새를 보인다. 특이한 점은 케이블 채널의 장르 선호 분포인데, 드라마와 VOD 채널에 시청량이 집중되어 있는 것을 볼 수 있다. 이로 미루어 보아 케이블에 대한 배경적 시청이 주로 드라마와 VOD 채널에서 이루어진다는 것을 짐작해볼 수 있다.

2.4. 케이블 중심 탐색형

'케이블 중심 탐색형' 가구 유형은 주로 케이블-습관적 세션 유형과 다중-탐색적 세션 유형을 중심으로 각 세션 유형이 조금씩 조합되어 있는데, 실제 채널 이용 비율을 살펴보면 케이블 채널에 상당히 집중되어 있는 모양새를 띤다. 즉, 이 시청 가구 유형은 다양한 채널을 탐색해보긴 하지만 결과적으로 케이블 채널로 이용이 수렴된다고 볼 수 있다. 케이블 채널의 장르 선호 분포는 앞서 살펴봤던 '케이블 중심 다식형' 가구 유형과 비슷한 모습을 보인다. 다만 '케이블 중심 다식형'이 VOD 채널에 대한 이용량도 상당히 있었던 것과는 대조적으로, '케이블 중심 탐

색형’ 가구 유형은 VOD 채널을 거의 이용하지 않은 것으로 나타났다.

2.5. 뉴스 중심 다식형

다음은 ‘뉴스 중심 다식형’의 시청 패턴에 대해 살펴본다. ‘뉴스 중심 다식형’은 도출된 모든 가구 유형에서 유일하게 세션 유형 조합 비율에서 종편-습관적 세션 유형의 구성비가 높은 편인데, 그럼에도 불구하고 채널의 이용량을 살펴보면 종합편성채널에 집중되어 있기 보다는 다양한 채널을 두루 이용하는 다식형 시청패턴을 보인다. 종합편성채널은 주로 뉴스를 중심으로 프로그램이 편성되어 있는데, ‘뉴스 중심 다식형’의 케이블 장르 선호 패턴에서도 케이블 뉴스 채널이 높은 것으로 보아 전반적으로 뉴스에 대한 선호가 높은 시청 가구 유형인 것으로 짐작해볼 수 있다. 특히 세션 유형이 주로 배경적 특성보다는 습관적 특성을 가지고 있는데, 이와 뉴스 시청 선호에 대한 관계를 살펴보면 기존 연구에서 밝혔듯 뉴스 콘텐츠 시청자 중 시청 가능한 시간대에 뉴스를 파악하고자 하거나 적극적인 시청자의 비율이 수동적이고 배경적으로 채널을 선택하는 시청자의 비율보다 높다는 사실과 어느 정도 일치한다고 여겨질 수 있다 [42].

2.6. 지상파 중심 다식형

주로 지상파-습관적 세션 유형으로 조합되어 있는 ‘지상파 중심 다식형’은 채널 시청량 에서도 지상파 채널에 대한 집중이 강하게 일어난

다. 대조적으로 종합편성채널에 대한 시청량은 거의 없는 편이며, 케이블에 대한 시청량이 두 번째를 차지한다. 이는 다중-탐색적 세션 유형과 같은 시청 행동으로 인해 케이블 채널에 대한 탐색과 시청이 일어나기 때문이라 생각해볼 수 있다. 케이블 채널 내에서의 장르 선호 분포는 드라마와 엔터테인먼트의 시청량이 타 장르에 비해 많은 모습을 보인다. 오락 및 드라마 장르를 주로 여성이 선호한다는 연구결과 [43]를 고려해보면, 이 유형의 가구는 주로 여성 가구원에게 채널 선택권이 있는 가구일 수 있음을 짐작해볼 수 있을 것이다.

2.7. 균형적 다식형

'균형적 다식형' 가구 유형은 타 가구 유형과 비교해 봤을 때 매우 독특한 세션 유형 분포를 가지는데, 모든 세션 유형이 고루 분포하는 모습을 보인다. 채널 시청량에 있어서도 종합편성채널에서의 시청량이 비교적 작긴 하지만, 각 카테고리의 채널을 두루 이용하는 모습을 보여 균형 있는 채널 이용 패턴을 보인다. 다만 케이블 채널에서 장르 선호 분포까지 균형 잡힌 모습은 아닌데, 드라마와 엔터테인먼트, 그리고 스포츠와 뉴스 순서대로 높은 시청량을 나타낸다.

2.8. 케이블 중심 습관적 편식형

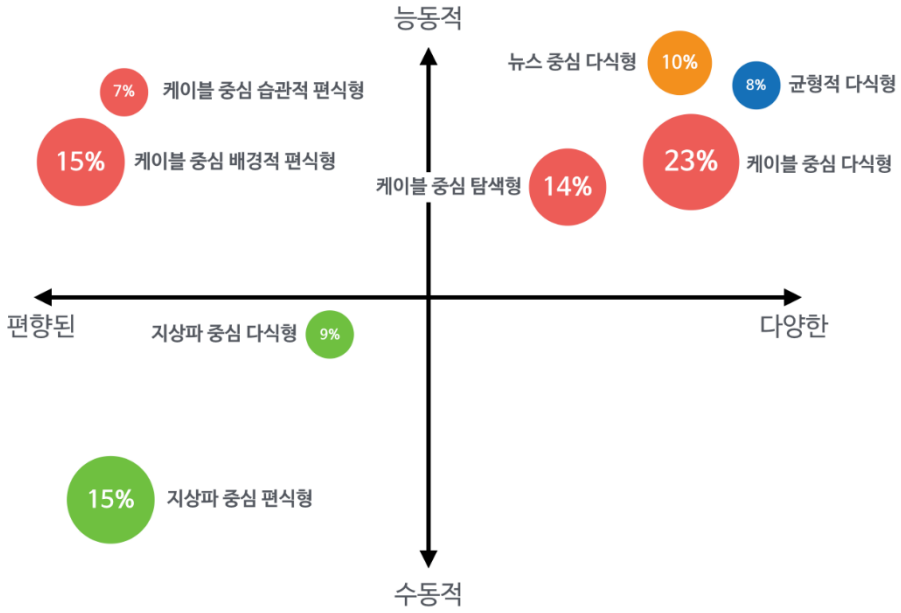
'케이블 중심 습관적 편식형' 가구 유형은 전체 세션의 7%를 차지한다. 이 가구 유형은 앞서 살펴본 '케이블 중심 배경적 편식형'과 상호

비교가 가능한데, ‘케이블 중심 배경적 편식형’이 케이블-배경적 세션 유형을 비교적 많이 보유하고 있는 것과 대조적으로 이 유형은 케이블-습관적 세션 유형이 다수를 차지하고 있다. 이에 따라 케이블 채널로 집중된 채널 이용량 분포를 보이는데, 흥미로운 점은 이 ‘케이블 중심 습관적 편식형’의 케이블 채널 장르 선호 경향이 ‘케이블 중심 배경적 편식형’의 그것과 판이하게 다르다는 것이다. 이를 통해 시청 행동 유형의 차이가 채널 이용 특성의 차이로 드러날 수 있다는 것을 실제로 확인할 수 있다. 케이블-습관적 행동 패턴과 케이블-배경적 행동 패턴의 구성비 차이에 따라 케이블 채널 장르의 이용 비율이 달라진다는 것을 통해, 단순히 장르 또는 채널의 이용량 및 빈도 등으로 시청자 혹은 시청 가구를 분석할 때보다 구체적인 행동 패턴에 대한 분석이 추가될 때 시청 가구에 대해 더욱 구체적으로 이해할 수 있는 실마리를 제공할 수 있다는 점을 확인할 수 있다.

2.9. 결과 정리

각 TV 시청 가구가 발생시킨 세션을 앞 절의 세션 유형으로 분류한 후 이를 토대로 클러스터링하여, 총 여덟 개의 시청 가구 클러스터를 구분해 유형화를 진행하였다. 각 시청 가구 유형은 각기 특징적인 세션 유형 구성비를 통해 TV 시청행태를 보였다. 시청 가구 유형의 세션 유형 구성비의 특징적인 부분으로는 균형적 다식형을 제외하고는 특정 채널 카테고리를 중심으로 강한 선호를 보이며 타 카테고리에 대해서 다식

적 혹은 편식적 시청 패턴을 보였다는 것이다. 균형적인 시청 패턴을 보인 그룹보다는 한 카테고리에 치우친 시청 패턴을 보인 그룹의 수가 더 많았다. 또한 케이블 채널 또는 뉴스 장르에 대한 선호를 보이는 몇몇 가구 유형의 경우 기존 연구에서 드러난 케이블 TV 또는 뉴스 프로그램 시청 경향에 대해 일정 부분 재확인해볼 수 있었다. 이에 따라, 각 가구 유형의 TV 시청 패턴을 소비의 다양성 측면에서 구별해볼 수 있다. 케이블 및 지상파를 중심으로 한 편식형 유형은 소비의 다양성 면에서 비교적 경직되어 있어 편향적인 콘텐츠 소비 패턴을 보인다. 지상파 중심 다식형의 경우 지상파 중심 편식형과 비교해 보았을 때 지상파에 덜 쏠려 있는 모양새를 보이긴 하나, 전반적으로 지상파에 높은 비중을 두고 있어 여전히 일정 정도의 편향성을 보인다. 이는 세션 유형의 구성 요소에 의한 차이로 보이는데, 지상파 중심 편식형 가구 유형이 월등히 높은 지상파-배경적 세션 유형을 가지고 있어 이러한 편향성의 차이가 나타난다고 할 수 있다.



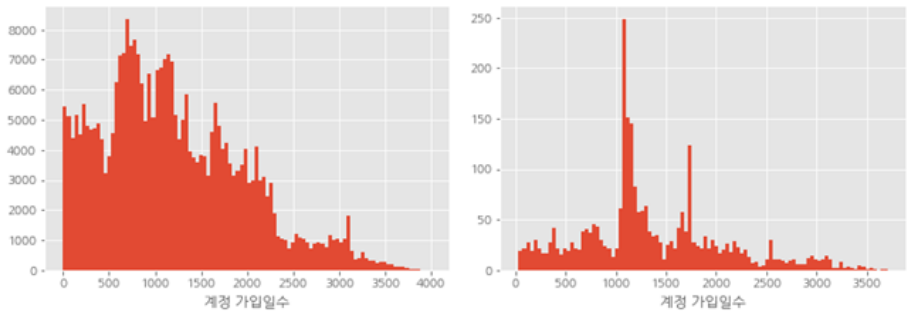
[그림 19] - 소비 다양성과 능동성에 따른 가구 유형 구분

콘텐츠 소비 다양성의 측면과 함께, 각 가구 유형이 주로 시청하는 채널의 카테고리(지상파·종편·케이블)에 따라 세션 유형과 마찬가지로 각 가구 유형의 선호 채널에 따른 능동성 정도를 파악할 수 있다. 결과적으로 [그림 19]와 같이 각 가구 유형의 콘텐츠 소비에 대한 특징을 소비의 다양성-편향성, 그리고 선호 충족의 능동성-수동성을 축으로 하여 나타냈다. 그림에서 볼 수 있듯, 지상파 위주의 시청 패턴을 보이는 가구 유형을 정도의 차이는 있지만 수동적이며 편향적인 유형으로 분류하였다. 그리고 선호에 따라 능동적 채널 선택을 보이지만 케이블 채널을 편향적으로 이용하는 케이블 중심 습관적·배경적 편식형 가구 유형을 능동적이며 편향적인 유형으로 분류하였다. 마지막으로 다식형 유형과 케

이블 중심 탐색형 유형을 능동적이며 다양한 소비를 행하는 유형으로 분류하였다.

제 3 절 세션 유형과 서비스 해지의 연관성

본 연구에서는 세션 유형과 서비스 해지 사이의 연관성을 밝히기 위해, 주어진 서비스 해지 데이터를 사용하여 각 시청 가구 유형 그룹의 해지율을 각각 계산하고 이와 각 시청 가구 유형의 세션 유형 구성비 사이의 피어슨 상관계수(Pearson Correlation Coefficient)를 구하였다. 이를 위해 먼저 서비스 해지자에 대한 전처리가 필요한데, 전체 가입가구에 대한 계정 가입일 수와 서비스 해지 가구에 대한 계정 가입일 수를 히스토그램으로 비교해보았다([그림 20]). 그래프를 살펴보면 전체 가입자와 서비스 해지 가구의 가입일 수 분포가 다름을 알 수 있는데, 특히 서비스 해지 가구의 경우 가입일 기준 1,090일 부근에서 피크가 형성되는 것을 볼 수 있다. 이는 주로 디지털 케이블 TV 서비스 가입 시 3년 약정 프로그램을 사용하는 것에서 비롯된 것으로 풀이된다. 3년 약정 기간이 종료되는 경우 별도 위약금 없이 서비스를 해지할 수 있고 이 때 타 케이블 TV 서비스 혹은 IPTV 서비스에서 전환 가입을 권유 받거나,



[그림 20] – 전체 가입가구 및 서비스 해지 가구의 계정 가입일수 분포

혹은 단순히 서비스에 대한 불만족으로 더 이상 서비스 구독을 유지하지 않아도 되기 때문이다.

따라서 서비스 구독 기간이 서비스 해지에 미치는 영향이 상당히 크기 때문에, 본 연구에서는 3년(1,095일)의 전후 1개월 간을 3년 약정 프로그램에 대한 조정 기간으로 보고, 따라서 전체 해지 데이터를 약정

프로그램 해지와 관련이 적은 그룹 및 많은 그룹으로 분할해 각각 분석을 진행하였다. 즉, 해지 가구 중 서비스 구독 기간이 1,065일 미만이거나 1,125일 초과인 서비스 해지 가구를 약정 해지와 관련이 적은 그룹, 서비스 구독 기간이 1,065일 이상 1,125일 이하인 해지 가구를 약정 해지와 관련이 큰 그룹으로 보았다. 주어진 해지 데이터 기간 동안 서비스 해지 가구는 총 2,672 가구로 전체 분석 가 구인 257,396 가구 중 1.04%를 차지하였으며, 이 중 약정 해지와 관련이 큰 그룹은 184 가구, 관련이 적은 그룹은 2,488 가구이다.

시청 가구 유형	시청 가구 수	해지 가구 수	해지율(%)
케이블 중심 다식형	56,567	410	0.72
지상파 중심 편식형	37,223	173	0.46
케이블 중심 배경적 편식형	36,790	315	0.86
케이블 중심 탐색형	35,066	276	0.79
뉴스 중심 다식형	24,144	138	0.57
지상파 중심 다식형	21,362	131	0.61
균형적 다식형	19,008	136	0.72
케이블 중심 습관적 편식형	16,820	126	0.75

[표 9] – 약정 해지 비관련 그룹의 각 시청 가구 유형별 해지율

[표 9]는 약정 해지와 관련이 적은 그룹 내의 각 시청 가구의 유형 그룹에 대한 해지율을 나타낸 것이다. 해지율을 살펴보면 지상파 중심 편식형과 뉴스 중심 다식형, 지상파 중심 다식형 유형이 낮은 해지율을 나타내는 것을 볼 수 있으며, 케이블 중심 배경적 편식형과 케이블 중심 습관적 편식형 가구 유형의 해지율이 비교적 높게 나타난다. 이에 따라 주로 시청하는 채널이 지상파일 경우 해지율이 낮으며, 반대로 케이블을 주로 시청하는 가구일수록 서비스를 해지할 가능성이 더 높다고 해석할 수 있다.

더욱 구체적으로 각 시청 행동이 서비스 해지와 연관되어 있는 정도에 대해 알아보기 위해, 각 시청 가구 유형의 각 세션 유형 구성비와 그룹 내 해지율 간의 상관계수를 구하였다. 각 세션 유형 구성비와 해지율 간 상관계수 결과는 [표 10]과 같다.

세션 유형	해지율과의 상관계수
케이블-배경적	**0.62**
케이블-습관적	0.57
지상파-배경적	**-0.81**
지상파-습관적	-0.56
종편-습관적	-0.42
다중-탐색적	0.36
다중-긴 탐색적	0.02

[표 10] - 약정 해지 비관련 그룹의 세션 유형 구성비와 해지율 간 상관계수

결과 표를 살펴보면, 대부분의 세션 유형이 해지율과 의미 있는 상관관계를 보인다는 것을 알 수 있다. 주로 지상파와 종합편성채널 위주로 시청하는 세션 유형의 구성비는 해지율과 음의 상관관계를 보이며, 케이블을 중심으로 시청하는 세션 유형의 구성비는 해지율과 양의 상관관계를 보인다. 또한 여러 카테고리의 채널을 탐색하는 다중-탐색적 세션 유형은 약한 양의 상관관계를 보이는 반면, 비슷한 속성이며 더 긴 평균 세션 길이를 가지는 다중-긴 탐색적 유형은 서비스 해지와 별다른 연관성이 없는 것으로 나타났다.

결과를 좀 더 자세히 구분해보면, 우선 각 세션이 주로 집중적으로 시청하는 채널의 카테고리가 해지와 큰 연관이 있다. 주로 케이블을 많이 시청하는 행동을 보일수록 지상파나 종합편성채널에 비해 자신의 선호나 취향에 따라 더욱 능동적으로 TV를 시청하는 것으로 해석할 수 있다. 따라서, 능동적인 시청 행위를 보일수록 서비스가 자신의 취향에 맞지 않는 것에 대해 더욱 강한 불만족을 느끼고 서비스를 해지할 가능성이 높은 것으로 보인다.

그러나 서비스 해지율과의 관계를 봤을 때 선호에 대한 능동성과 채널 탐색에 대한 적극성은 그리 큰 연관을 보이지는 않는 것으로 짐작된다. 다중-탐색적 세션 유형 및 다중-긴 탐색적 세션 유형은 타 세션 유형에 비해 더욱 적극적으로 채널을 탐색하고 서비스를 이용하는 행동 유형인데, 이 두 세션 유형은 오히려 서비스 해지와 약한 양의 상관관계

를 갖거나 상관관계가 극히 미미하다. 따라서, 서비스 해지와 관련 있는 시청자의 성향은 적극성보다는 자신의 선호에 따라 채널을 선택해 시청하는 능동성이라고 볼 수 있다.

마지막으로 지상파와 종합편성채널의 경우를 살펴보면, 모든 세션 유형의 구성비가 서비스 해지율과 음의 상관관계를 지닌다. 특히 지상파-배경적 세션 유형은 서비스 해지에 대해 매우 강한 음의 상관관계를 가진다. 이는 이러한 유형의 행동을 주로 보이는 시청자들이 전통적 관점에서의 수동적 시청자를 대변해 굳이 적극적으로 서비스를 해지하고 타 서비스를 가입할 필요를 느끼지 못하는 경우로 해석할 수도 있으며, 혹은 식당이나 병원 등의 사업장에서 디지털 케이블 TV를 도입하고 배경적으로 사용하는 곳이므로 빈번한 서비스 전환이 잘 일어나지 않는 환경이라 짐작할 수도 있다. 이외에도 시청 패턴의 행동적 특성에 따른 차이로 인해 지상파-습관적 세션 유형은 서비스 해지와 비교적 뚜렷한 음의 상관관계를 가진다. 이는 역시 지상파 채널을 위주로 시청하는 시청 가구 집단의 수동성에서 비롯된 것이라 짐작 가능하며, 상관계수의 비교를 통해 같은 지상파 선호 시청 패턴이라도 배경적인 시청 패턴과 그렇지 않은 시청 패턴의 차이, 즉 시청 행동의 적극성이 일정 부분 서비스 이용에 있어 차이를 보인다는 것을 보여준다. 종편-습관적 세션 유형 또한 서비스 해지에 대해 뚜렷한 음의 상관관계를 가져, 지상파-습관적 세션 유형과 서비스 해지에 있어 비슷한 특징을 가진다는 것을 알 수 있다.

다음으로 약정 해지와 직접적인 관련이 큰 그룹 내의 각 시청 가구 유형별 해지율과 세션 유형 구성비 간의 연관성을 알아보았다. [표 11]은 이 그룹의 각 시청 가구 유형별 해지율을 나타낸 것이다. 약정 해지와 관련이 적은 그룹에서는 케이블 중심 배경적 편식형의 해지율이 가장 높았던 것과 달리, 이 그룹에서는 케이블 중심 다식형과 함께 균형적 다식형 가구 유형의 해지율이 높은 것을 볼 수 있다. 즉, 이는 세션 유형 구성비와 해지율 간의 상관관계가 앞의 그룹과 다르게 도출될 수 있음을 의미한다.

시청 가구 유형	시청 가구 수	해지 가구 수	해지율(%)
케이블 중심 다식형	2,004	53	2.64
지상파 중심 편식형	1,233	24	1.95
케이블 중심 배경적 편식형	1,326	27	2.04
케이블 중심 탐색형	1,131	27	2.39
뉴스 중심 다식형	825	16	1.94
지상파 중심 다식형	679	10	1.47
균형적 다식형	601	17	2.83
케이블 중심 습관적 편식형	550	10	1.82

[표 11] - 약정 해지 관련 그룹의 각 시청 가구 유형별 해지율

다음은 앞의 그룹과 마찬가지로, 약정 해지 관련 그룹의 각 시청 가구 유형에 대해 세션 유형 구성비와 유형 내 해지율 간의 상관관계를 계산하였다. 각 세션 유형 구성비와 해지율 간의 상관관계는 [표 12]와 같다.

세션 유형	해지율과의 상관계수
케이블-배경적	0.12
케이블-습관적	-0.19
지상파-배경적	0.01
지상파-습관적	**-0.48**
종편-습관적	-0.07
다중-탐색적	0.41
다중-긴 탐색적	**0.65**

[표 12] - 약정 해지 관련 그룹의 세션 유형 구성비와 해지율 간 상관계수

표를 보면 앞서 약정 해지와 관련이 적은 그룹에서 도출된 결과와 그 값이 다른 것을 알 수 있다. 3년 약정의 영향력을 줄인 그룹에서는 지상파-배경적 세션 유형의 구성비가 해지율에 대해 매우 큰 음의 상관관계를 가졌는데, 이 그룹에서는 거의 상관관계가 없는 것으로 나타났다. 또한, 케이블-배경적 세션 유형의 상관계수가 0.61에서 0.12로 감소했다. 케이블-습관적 세션 유형의 상관계수는 0.58에서 -0.19로 변화해, 상관관계가 음의 방향으로 변화되었다. 특히 다중-긴 탐색적 세션 유형은 앞 그룹에서 거의 상관관계가 없는 것으로 나타났는데, 이 그룹에서는 0.65로 비교적 높은 양의 상관관계를 보였다.

결과를 분석해 보면, 많은 수의 세션 유형이 실질적으로 해지율에 큰 영향을 미치지는 않는 것으로 나타났다. 지상파-배경적 및 종편-습관적 유형은 해지율과 큰 관련이 없어, 이러한 유형의 행동은 많고 적음과 관계 없이 3년 약정 만기가 도래하면 서비스를 해지할 수 있는 것으로 보인다.

3년 약정 프로그램이 종료되었을 때 가장 영향을 많이 받는 그룹은 다중-긴 탐색적 행동을 많이 보이는 시청 가구들이다. 여러 카테고리의 채널을 구분 없이, 긴 시간에 걸쳐 탐색하고 프로그램을 시청하는 행동이 많다는 것은, 케이블 채널이나 지상파 또는 종편 채널에 집중하는 시청 패턴에 비해 이들이 서비스 제공자 관점에서 부동층과 같은 그룹임을 시사한다. 즉, 자신의 프로그램 장르에 대한 선호가 확고해 케이블 채널 위주로 TV를 시청하거나 지상파 및 종합편성채널을 애호하는 기존의 수동적이며 전통적인 시청 가구 집단이 아닌, 특별한 선호 없이 여가 시간을 보내기 위해 여러 채널과 프로그램을 방랑하는 시청 패턴을 보인다고 볼 수 있다. 따라서, 3년 약정 프로그램의 만기가 다가왔을 때에도 제공되는 서비스에 대한 선호에 따라 서비스 이용을 지속하거나, 수동적인 시청 패턴으로 인해 서비스 이용을 해지하지 않는 것과는 다르게, 타 경쟁 서비스의 프로모션 등에 대해 반응을 보이고 타 서비스로 이전할 가능성이 다른 행동 유형이 지배적인 그룹에 비해 높다고 볼 수 있다. 특히 두 가지 탐색적 유형이 모두 양의 상관관계를 보이면서 그 시간이 길수록 상관관계가 더 커지는 것은, 이러한 행동 유형의 부동층적 속성이 시간에 따라 달라지기 때문이라 해석해볼 수 있을 것이다.

제 6 장 결 론

제 1 절 연구 요약

본 연구는 기존 TV 시청자 연구 및 분석 방식이 비용 및 확장성에 대해 지닌 한계를 극복함과 동시에 미시적 시청 행동에 대한 연구와 거시적 채널 소비에 대한 연구의 간극을 메우고자 디지털 케이블 TV의 셋톱박스를 통한 대규모 TV 시청 로그를 바탕으로 TV 시청 가구를 시청 행동을 기반으로 분석하였다. 특히 TV 시청 가구의 시청 행태에 대한 보다 구체적인 이해를 돕기 위해 웹 사용 마이닝 분야의 행동 기반 클러스터링 방법론을 기반으로 한 시청 행동 및 시청 가구 클러스터링 기법을 제안하였다. 또한 제안된 방안을 통해 시청 행동을 세션화하고 특질을 추출한 후 클러스터링하여 일곱 가지의 세션 유형을 도출하였으며, 이 세션 유형의 구성비로 시청 가구를 묘사하여 클러스터링함으로써 여덟 가지의 시청 가구 유형을 도출하였다. 마지막으로는 이러한 행동 기반 분석 프레임워크를 통해 도출된 시청 행위 유형 및 가구 유형의 실질적 유효성을 검증해보기 위하여 디지털 케이블 TV의 서비스 해지와 시청 행위 유형 간의 연관성에 대해 알아보았다.

연구 결과 도출된 세션 유형은 전체 세션에 대한 구성비 순서대로 케이블-배경적, 케이블-습관적, 지상파-배경적, 지상파-습관적, 다중-탐

색적, 종편-습관적, 다중-긴 탐색적 유형이며, 이를 통해 전체 시청 가구에 대한 구성비 순서대로 케이블 중심 다식형, 지상파 중심 편식형, 케이블 중심 배경적 편식형, 케이블 중심 탐색형, 뉴스 중심 다식형, 지상파 중심 다식형, 균형적 다식형, 그리고 케이블 중심 습관적 편식형 유형이 시청 가구 유형으로 도출되었다. 또한 각 세션 유형과 서비스 해지 간의 연관성을 파악하기 위해 진행된 연구에서는 채널 소비의 능동성과 적극성에 따라 세션 유형이 서비스 해지율에 대해 보이는 상관관계에 차이가 있음을 입증하였다.

각 세션 유형은 주로 어떤 카테고리의 채널을 봤는지와 함께 소수의 채널을 옮겨 다니는 습관적 행태, 한 채널을 오래 시청하다 종료하는 배경적 행태, 그리고 카테고리 구분 없이 여러 채널을 탐색하는 탐색적 행태를 기준으로 분류되었다. 이를 콘텐츠 소비에 대한 능동성 및 채널 탐색에 대한 적극성을 기준으로 다시 분류하여 TV 시청 시작부터 종료에 이르는 개별 시청 행위들이 어떤 속성의 패턴을 보이는지에 대해 분석해보았다. 이 세션 유형을 통해 도출된 시청 가구 유형을 통해 시청 가구들이 특정 카테고리나 장르를 기반으로 강한 선호를 보이거나 균형적인 시청 패턴을 보이는 모습을 관찰할 수 있었으며, 이를 콘텐츠 소비에 대한 능동성과 소비의 다양성 측면에서 다시 분류하여 이해를 도모하였다.

제 2 절 연구의 시사점

과거의 단순하고 일방적인 TV 시청 환경에서 벗어나, TV의 다채널화와 함께 다양한 형태의 매체가 등장해 콘텐츠 소비 행태가 과거에 비해 매우 복잡해졌다. 이에 따라 TV 시청 행태에 영향을 미치는 요인이 늘어나면서 기존 연구 방식이 채택했던 전통적인 인구통계학적 정보, 교육 및 소득수준에 따른 요인만으로는 복잡한 TV 시청 행태와 그 행태를 보이는 시청자에 대한 이해가 힘들어졌다.

이에 시청자의 TV 시청 행동 자체로 돌아가 행동 유형을 분류하고 분석하는 사용자 조사를 기반으로 TV 시청 행태를 이해하려는 노력이 이어졌다. 많은 경우 이러한 방식의 사용자 조사는 잘 설계된 연구를 바탕으로 풍부한 데이터를 통해 다양한 방식으로 TV 시청자 및 시청 행위에 대한 이해를 도왔다. 하지만 연구의 설계 및 사용자의 모집 등에 대한 확장성의 한계와 함께 오랜 시간에 걸친 종단 연구가 어렵다는 한계점을 내포하고 있다.

이와 비교하여 대규모 셋톱박스 로그를 활용하여 데이터를 기반으로 TV 시청 데이터를 분석하고 시청 행위 및 시청 가구를 유형화하는 방식은 연구자에게 추가적인 데이터 획득을 위한 노력이 불필요하고, 전수 데이터를 사용하므로 서비스 이용자에 대한 대표성을 추가적으로 획득할 수 있다는 장점을 가지고 있다. 그리고 이러한 데이터 기반 탐색적

연구의 또다른 장점으로는 기존에 알려지지 않은 행동 패턴 또한 도출해 낼 수 있다는 점이 있다. 이를 통해 새롭게 발굴된 행동 패턴 자체에 초점을 맞춰 시청자들이 어떤 이유와 요인으로 특정 행동 패턴을 일으키는지에 대해 드릴다운 분석을 통해 더욱 깊은 이해를 도모하는 데에 기여할 수 있다고 기대한다.

본 연구는 대규모 데이터를 이용한 데이터 분석적 접근법을 활용하여 현재의 TV 시청 환경에서 TV 시청 행동 및 시청 가구에 대한 이해의 폭을 넓힘으로써, 기존 연구의 결과와 맥을 잇는 동시에 기존 연구에서 보이지 않았던 추가적인 시청 패턴을 발굴해낼 수 있었다. 또한 다채널 환경에서의 TV 시청 행위에 대해 이루어진 기존의 연구는 주로 거시적인 관점에서 각 시청자가 보이는 채널 소비 행태를 채널 레퍼토리 등을 통해 이해하고자 하였는데, 결국 이러한 채널 레퍼토리는 시청을 시작하고 종료하기까지의 개별 시청 행동의 결과물로써 이루어진 것이다. 따라서, 본 연구에서 제안한 분석 방법과 그 연구 결과를 통해 기존 채널 레퍼토리 등의 연구를 더욱 확장시켜, 구체적으로 어떠한 개별적 시청 행위들을 통해 채널 레퍼토리가 형성되거나 콘텐츠 소비에 대한 다양성 혹은 분극화 등이 일어나는지에 대해 더욱 폭넓은 TV 시청자에 대한 이해를 가능케 할 수 있다는 데에 학술적인 의의가 있다.

또한 본 연구는 실제 디지털 케이블 TV 서비스의 대규모 셋톱박스 시청 로그를 통해 실질적인 빅데이터 분석을 실시하여, 실제 시청 가구

의 시청 행위에 대해 분석하였다. 이를 통해 특정 가입 지역의 전수 데이터를 바탕으로 해당 지역의 서비스 이용자에 대한 대표성을 획득할 수 있었다. 또한 본 연구에서 제안한 분석 프레임워크의 결과물로서 얻어진 시청 행위 유형과 시청 가구 유형이 실제로 서비스 해지율에 대해 의미 있는 연관성을 보였다. 따라서 연구 결과를 바탕으로 실제 서비스 내에서 가입 가구를 세분화하고 유목화 하여 서비스 마케팅이나 맞춤형 개인화 등에 본 연구의 연구 결과를 활용하거나, 개별 가입가구의 시청 행위 유형 조합에 따라 세부적으로 요금제 및 약정기간 등을 조정하는 데에 활용할 수 있다는 데에서 산업적인 의의가 있다.

마지막으로 본 연구는 TV 시청자의 시청 패턴에 대해 연구 및 분석을 실시하였는데, TV 시청 행태 및 시청자 프로파일은 경영학, 컴퓨터공학, 산업공학 등 다양한 학문 분야의 관심사가 되어 왔다. 이를 통해 본 연구의 주제가 융합적인 연구 대상임을 알 수 있으며, 또한 기존의 타 연구 분야인 웹 사용 마이닝에서의 클릭스트림 및 세션 기반 행위 분석 방법론을 차용하여 TV 시청자의 시청 행위를 대상으로 분석을 실시하여, 주제 및 방법론에 있어 본 연구에 융합적인 의의가 있다고 할 수 있다.

제 3 절 연구의 한계

본 연구에서 제안한 시청 행위 및 시청 가구의 유형화에 따른 분석은 TV 시청 행위에 대한 더욱 세밀한 이해를 가능하게 하지만, 여러 한계점을 가진다. 먼저, 본 연구에서 사용한 데이터는 시청 로그의 수집 단위가 시청자 개개인이 아닌 시청 가구다. 그리고 데이터의 한계로 인해 서비스 가입가구의 가구원 수, 혹은 가입자의 성별 및 나이 등에 대한 인구통계학적 정보를 배제한 채로 분석을 진행하였다. 따라서 개인과 가구에 따른 차이, 가구원 수에 의한 차이, 성별 및 나이 등의 요인에 의한 영향을 고려하지 못했다는 한계가 있다.

또한 본 연구에 사용된 데이터는 CJ 헬로tv 전체 가입자 중 특정 지역 가입자의 데이터에 한정되었다. 해당 지역 유료 방송 서비스 중 하나의 서비스에서 제공된 데이터로 연구 및 분석을 진행했으므로, 지역 전체, 나아가 국내 시청자 전체에 대한 일반화를 할 수 없으며, 단지 전체 시청 가구 중 일부 부분집합에 대해 시청 행위의 경향을 파악해보았다는 한계가 있다.

마지막으로 본 연구에서는 시청 로그를 변환한 세션에서 특질을 추출하거나 세션 유형, 혹은 가구 유형에 대한 분석을 진행할 때 시청 행위가 일어난 요일이나 시간대에 대해 고려하지 않았다. TV 시청 행위는 요일이나 시간대에 따라 다른 특징을 보일 수 있어 시간적인 요인에 따

른 영향을 무시할 수 없으나, 본 연구에서는 문제를 단순화하고 도출된 유형에 대한 이해를 시도함에 있어 시간적 요인을 고려하지 않았다. 따라서 도출된 각 유형 및 그에 대한 분석에 있어 한계점이 명확하다. 향후 연구를 지속함에 있어 시간적인 요인에 대한 적극적인 고려가 반드시 필요하다고 할 수 있다.

참고문헌

[1] 방송통신위원회, “방송매체 이용행태조사,” 2016.

[2] 심미선, “다매체 시대 미디어 레퍼토리 유형에 관한 연구,” *한국방송학보,* 제 21, p. 351–390, 2007.

[3] B. A. Bondad-Brown, R. E. Rice and K. E. Pearce, "Influences on TV viewing and online user-shared video use: Demographics, generations, contextual age, media use, motivations, and audience activity," *Journal of Broadcasting & Electronic Media,* vol. 56, no. 4, pp. 471-493, 2012.

[4] C. Wallis, "The Multitasking Generation," *Time Magazine,* vol. 167, no. 13, pp. 48-55, 2006.

[5] W. E. Spangler, M. Gal-Or and J. H. May, "Using data mining to profile TV viewers," *Communications of the ACM,* vol. 46, no. 12, pp. 66-72, 2003.

[6] Z. Yu and X. Zhou, "TV3P: an adaptive assistant for personalized TV," *IEEE Transactions on Consumer Electronics,* vol. 50, no. 1, pp. 393-399, 2004.

[7] D. Goren-Bar and O. Glinansky, "FIT-recommend ing TV programs to family members," *Computers & Graphics,* vol. 28, no. 2, pp. 149-156, 2004.

[8] 심미선, 김은미 그리고 이준웅, “라이프 스타일에 따른 텔레비전 시청패턴 연구,” *한국언론학보,* 제 48, 번호: 2, pp. 189-217, 2004.

[9] A. M. Rubin, "Television uses and gratifications: The interactions of viewing patterns and motivations," *Journal of Broadcasting & Electronic Media,* vol. 27, no. 1, pp. 37-51, 1983.

[10] R. Warren, "Parental mediation of children's television viewing in low-income families," *Journal of Communication,* vol. 55, no. 4, pp. 847-863, 2005.

[11] H. D. Rozanski, G. Bollman and M. Lipman, "Seize the occasion! The seven-segment system for online marketing," *Strategy and Business,* pp. 42-53, 2001.

[12] J. Vanattenhoven and D. Geerts, "Second-screen use in the home: an ethnographic study," *Bridging people, places & platforms: Proceedings EuroITV2012,* pp. 162-173, 2012.

[13] W. Trouleau, A. Ashkan, W. Ding and B. Eriksson, "Just One More: Modeling Binge Watching Behavior," *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp. 1215-1224, 2016.

[14] N. F. Krueger, M. D. Reilly and A. L. Carsrud, "Competing models of entrepreneurial intentions," *Journal of business venturing,* vol. 15, no. 5, pp. 411-432, 2000.

[15] 강남준 그리고 김은미, “다중 미디어 이용의 측정과 개념화: 오디언스를 향한 새로운 시선,” *언론정보연구,* 제 47, 번호: 2, pp. 5-39, 2010.

[16] 박인곤 그리고 강형구, “디지털 케이블 TV 수용자의 이용과 충족, 디지털 케이블 수용 여부에 대한 잠재적 변수의 영향에 관한 연구,” *한국방송학보,* 제 21, 번호: 6, pp. 156-192, 2007.

[17] 이상식 그리고 김관규, “케이블 TV 의 채널 레퍼토리(repertoire)에 관한 연구,” *한국언론학보,* 제 45, p. 268–297, 2001.

[18] L. Lu, M. Dunham and Y. Meng, "Mining significant usage patterns from clickstream data," *In Proc. of WebKDD,* 2005.

[19] G. Wang, T. Konolige, C. Wilson, X. Wang, H. Zheng and B. Y. Zhao, "You Are How You Click: Clickstream Analysis for Sybil Detection," *In Proc. of USENIX Security,* 2013b.

[20] M. Spence and B. Owen, "Television programming, monopolistic competition, and," *The Quarterly Journal of Economics,* pp. 103-126, 1977.

[21] D. Gensch and P. Shaman, "Models of Competitive Television Ratings," *Journal of Marketing Research,* vol. 17, no. 3, pp. 307-315, 1980.

[22] J. G. Webster, "Program audience duplication: A study of television inheritance effects," *Journal of Broadcasting & Electronic Media,* vol. 29, pp. 121-133, 2015.

[23] 최수진, “케이블 TV 채널 소비와 문화적 다식성-편식성 가설에 대한 탐색,” *한국언론학보,* 제 57, 번호: 6, pp. 527-551, 2013.

[24] 배진아, “지상파 및 다채널 텔레비전 시청의 수동성과 능동성 비교 연구: 시청 패턴의 유사성 분석을 중심으로,” *한국언론학보,* 제 48, 번호: 5, pp. 30-52, 2004.

[25] P. M. Napoli, "Deconstructing the diversity principle," *Journal of Communication,* vol. 49, no. 4, pp. 7-34, 1999.

[26] C. Heeter, "Program selection with abundance of choice," *Human Communication Research,* vol. 12, no. 1, p. 126–152, 1985.

[27] 강남준 그리고 조성동, "다채널 시대 시청행태 변화에 대한 연구: 60 대 이상 노년층의 시청양상 변화를 중심으로," *미디어 경제와 문화,* 제 5, 번호: 4, pp. 7-45, 2007.

[28] 정영호 그리고 강남준, "네트워크 분석을 활용한 다채널 시대의 시청행태 분석," *한국방송학보,* 제 24, 번호: 6, pp. 323-363, 2010.

[29] J. G. Webster and T. B. Ksiazek, "The dynamics of audience fragmentation: Public attention in an age of digital media," *Journal of Communication,* vol. 62, no. 1, pp. 39-56, 2012.

[30] L. Barkhuus, "Television on the internet: new practices, new viewers," *CHI'09 Extended Abstracts on Human Factors in Computing Systems,* pp. 2479-2488, 2009.

[31] J. Lull, "Family communication patterns and the social uses of television," *Commun Res,* vol. 7, p. 319–333, 1980.

[32] D. Saxbe, A. Graesch and M. Alvik, "Television as a social or solo activity: understanding families' everyday television viewing patterns," *Commun Res Rep,* vol. 28, p. 180–189, 2011.

[33] K. Mercer, A. May and V. Mitchel, "Designing for video: investigating the contextual cues within viewing situations," *Pers Ubiquit Comput,* vol. 18, p. 723–735, 2014.

[34] J. Abreu, P. Almeida, B. Teles and M. Reis, "Viewer behaviors and practices in the (new) television environment," *Proceedings of the 11th European conference on interactive TV and video,* pp. 5-12, 2013.

[35] A. J. B. Chaney, M. Gartrell and J. Hofman, "A large-scale exploration of group viewing patterns," *Proceedings of the 2014 ACM international conference on interactive experiences for TV and online video,* p. 31–38, 2014.

[36] E. Adar, J. Teevan and S. T. Dumais, "Large Scale Analysis of Web Revisitation Patterns," *In Proc. of CHI,* 2008.

[37] H. Obendorf, H. Weinreich, E. Herder and M. Mayer, "Web page revisitation revisited: implications of a long-term click-stream study of browser usage," *In Proc. of CHI,* 2007.

[38] S. Gündüz and M. T. Özsu, "A Web page prediction model based on click-stream tree representation of user behavior," *In Proc. of SIGKDD,* 2003.

[39] S. Ghemawat, H. Gobioff and a. S.-T. Leung, "The Google file system," *In 19th Symposium on Operating Systems Principles,* pp. 29-43, 2003.

[40] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," *In OSDI'04: Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation,* pp. 10-10, 2004.

[41] M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M. J. Franklin, A. Ghodsi, J. Gonzalez and S. Shenker, "Apache Spark: A Unified

Engine for Big Data Processing," *Commun. ACM,* vol. 59, no. 11, pp. 56-65, 2016.

[42] 김승환, "TV 뉴스 콘텐츠의 채널 선택 유형에 따른 수용자 특성," *한국콘텐츠학회논문지,* 제 7, 번호: 6, pp. 99-106, 2007.

[43] 심미선, "융합매체환경 하에서의 장르이용에 관한 연구: 지상파와 케이블 텔레비전을 중심으로," *방송연구,* 제 겨울호, pp. 175-220, 2007.

[44] 심미선, "다중 미디어 이용 연구에 관한 비판적 고찰," *언론정보연구,* 제 47, pp. 40-73, 2010.

[45] 이현우 그리고 오형일, "지상파 콘텐츠 시청서비스 조합유형 분석," *한국방송학보,* 제 27, pp. 251-293, 2013.

[46] 강남준, 이종영 그리고 이혜미, "군집분석 방법을 사용한 미디어 레퍼토리 유형분석," *한국방송학보,* 제 22, p. 7–46, 2008.

[47] U. Hasebrink and J. Popp, "Media repertoires as a result of selective media use. A conceptual approach to the analysis of patterns of exposure," *Communications,* vol. 31, no. 3, p. 369–387, 2006.

Abstract

Analysis of Viewing Behavior and Household Type through Large-scale TV Log Clustering

Taeyoung Lee

Program in Digital Contents and Information Studies

Department of Transdisciplinary Studies

The Graduate School

Seoul National University

Recently, TV viewing behavior has become very complicated. They are interacting with various media and content supply services, and are displaying complex viewing behavior. The TV viewing environment has changed to a complex environment that is much harder to anticipate due to changes in the content platform and device environment.

Even when the environment surrounding the TV becomes more complicated, understanding of watching TV is still important. As the N-screen viewing environment becomes more common, the proportion of TVs is declining. However, people still spend a lot of time watching TV, and TV still plays an important role in content consumption. Even in a different environment, watching TV is still an important activity, suggesting that it is necessary to understand TV viewers and viewing behavior in a complex environment.

The purpose of this study is to overcome limitations such as scalability of existing researches which tried to understand TV watching based on behavior and ultimately to expand the understanding of TV

viewing in a diversified TV viewing environment. Based on the large-scale TV viewing logs obtained through the cable TV set-top box logs, we have proposed a framework for grouping the TV viewing patterns into behavior-based and using the patterns to understand TV viewers. To this end, we applied the session clustering-based segmentation technique, which was used in the Web Usage Mining field, to the TV viewing log. We also tried to verify the validity of our approach by looking at the correlation between typed viewing behavior and service cancellation rate.

Through the proposed analysis framework, we derived a total of 7 types of viewing behaviors and 8 types of household in combination. In addition, we can confirm that the type of viewing behaviors derived from this study can explain the service cancellation meaningfully through the correlation between the service cancellation rate and the viewing behavior type composition ratio in each type of household group.

By analyzing the viewing pattern based on the behavior using the proposed analysis framework, we can extend the previous research in the existing macro context to help a richer understanding of TV viewing behavior in the current media environment.

Keywords : TV Viewing Behavior Analysis, Big Data Analysis, Behavior-based Analysis

Student Number : 2015-26038